Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER THESIS

Predictability of security returns using Twitter sentiment

Author: **Bc. Marek Fremunt**

Supervisor: **PhDr. Jozef Baruník, Ph.D.**

Academic Year: **2014/2015**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 14, 2015

Signature

Acknowledgments

I am honestly grateful to my supervisor, Jozef Baruník, for his valuable suggestions, comments, support and willingness to answer all my questions. His guidance was more than necessary for the completion of this thesis. I, too, am very grateful to him for providing me with the market data. I would also like to express my gratitude to Jiří Cidlina for helping me with obtaining Twitter data and its processing.

Abstract

This work concentrates on exploring the influence of social networks to financial markets. We have introduced a novel approach to Twitter sentiment analysis, in which we collect continuous stream of data and analyze it. Our original data set contains over 200 million English written Tweets from the period between July 1, 2014 and October 9, 2014. Twitter sentiment is used as a good representative of investors' mood. On hourly data we investigate how investors are influenced by basic emotions, moods and sentiment in their decision making processes as well as the influence of keywords related to specific securities and FOREX symbols. Particularly, we examine the relationships between Twitter-based variables and returns as well as volatility of several financial instruments on a wide range of data including commodities, currencies and S&P 500 Cash Index. We show that Twitter sentiment influences volatility of securities' returns, tested and shown on both conditional and realized volatility models. We also describe the effect of Twitter sentiment on securities' returns. Moreover, we reveal the influence of basic emotions on investors' decision making processes. Our results suggest that investors are influenced by emotions and moods, especially at longer investment horizons. The impact of emotions at shorter investment horizons is limited and differs for particular securities as well as emotions.

JEL Classification	C52, C58, G11, G14, G17
Keywords	Twitter sentiment, security returns, volatility, Behavioral finance

Author's e-mail	`marek.fremunt@seznam.cz`
Supervisor's e-mail	`barunik@fsv.cuni.cz`

Abstrakt

Tato práce se zaměřuje na zkoumání vlivu sociálních sítí na finanční trhy. Náš nový přístup k analýze a měření nálady na Twitteru je použit na vzorku více než 200 milionů Tweetů z období od 1. července 2014 do 9. října 2014. Nálada na Twitteru je použita jako proxy proměnná pro náladu investorů. Na hodinové frekvenci dat zkoumáme vliv základních emocí, nálad a mínění na rozhodovací procesy investorů. Konkrétně, zkoumáme vztahy mezi proměnnými odvozenými z Twitterových dat a výnosy i volatilitou konkrétních finančních instrumentů, zahrnující např. komoditní a měnové futures. Vliv nálady na Twitteru na volatilitu výnosů finančních instrumentů testujeme jak na modelech podmíněné volatility, tak realizované volatility. Dále odhalujeme vliv základních emocí na rozhodování investorů. Naše výsledky naznačují, že investoři jsou nejvíce ovlivněni emocemi a náladou na delších investičních horizontech, zatímco v krátkém období je vliv emocí omezený.

Klasifikace JEL	C52, C58, G11, G14, G17
Klíčová slova	Nálada na Twitteru, výnosy finančních instrumentů, volatilita, Behaviorální finance

E-mail autora	`marek.fremunt@seznam.cz`
E-mail vedoucího práce	`barunik@fsv.cuni.cz`

Contents

List of Tables **viii**

List of Figures **x**

Acronyms **xii**

Thesis Proposal **xiii**

1 Introduction **1**

2 Literature review **4**

3 Methodology **9**

3.1 Hypotheses 10

3.2 Granger causality analysis 10

3.3 ARIMA models 11

3.4 Conditional Heteroscedastic Models 12

3.4.1 ARCH models 12

3.4.2 GARCH models 14

3.5 Realized Volatility 15

3.5.1 HAR models 15

3.5.2 HAR-RV model 15

3.6 Wavelets 17

3.6.1 Continuous Wavelet Transform 18

3.6.2 Wavelet Coherence 19

3.6.3 Wavelet coherence phase 20

4 Data description **21**

4.1 Twitter Sentiment tracking methods 21

4.2 Twitter Sentiment text analysis 22

4.3 Twitter Sentiment Extraction and Processing 22
4.4 Twitter Sentiment variables 23
4.5 Adjusting the time-frequencies of the Twitter sentiment and financial market data 25
4.6 Statistics of the Twitter variables 25
4.7 Financial Market Data 28
4.8 Statistics of the securities' returns 29
4.8.1 Japanese Yen (JY) 29
4.8.2 S&P 29
4.8.3 Natural Gas NYMEX (NG) 31
4.8.4 Silver COMEX (SV) 31
4.8.5 Light Crude NYMEX (CL) 32
4.8.6 Canadian Dollar (CD) 33

5 Empirical results 36
5.1 Granger causality analysis 36
5.2 Modelling returns 38
5.3 Volatility estimation 42
5.3.1 Conditional Volatility models 42
5.3.2 Realized Volatility models 49

6 Explanatory power of the variables using Wavelet Coherence 56
6.1 JY 56
6.2 S&P 57

7 Conclusion 63

Bibliography 68

A Tables I

List of Tables

4.1 Twitter sentiment variables 23
4.2 Statistics of the mood and tone variables 26
4.3 Statistics of the emotion variables 28
4.4 Statistics of JY 30
4.5 Statistics of S&P 31
4.6 Statistics of NG 32
4.7 Statistics of SV 33
4.8 Statistics of CL 34
4.9 Statistics of CD 35

5.1 GCA 38
5.2 GCA2 38
5.3 Modelling JY returns using AR(1) 40
5.4 Explanatory power of the hourly logarithmic differenced variables for S&P volatility 43
5.5 Explanatory power of the hourly logarithmic differenced variables for JPY volatility 44
5.6 Explanatory power of the GARCH(1,1) model for CD 46
5.7 Explanatory power of the GARCH(1,1) model for SV 46
5.8 Explanatory power of the GARCH(1,1) model for NG 48
5.9 VAR model - Estimation Results for S&P 50
5.10 HAR-RV model 53
5.11 HAR-RV model 54
5.12 HAR-RV model 55

A.1 Statistics of the emotion variables II I
A.2 Statistics of tag variables II II
A.3 Stationarity tests III
A.4 Stationarity tests IV

A.5 Stationarity tests V
A.6 Additional test statistics related to the explanatory power of the HAR-RV model VI
A.7 Additional test statistics related to the explanatory power of the HAR-RV model VII
A.8 Additional test statistics related to the explanatory power of the HAR-RV model VIII

List of Figures

2.1 Sentiment analysis from Medhat *et al.* (2014) 5

3.1 Plot of the Complex Morlet Wavelet with $f_0 = 6$ 19

4.1 Twitter sentiment - emotion variables based on Scott (2011) . . 24
4.2 Twitter sentiment - mood and tone variables based on Scott (2011) 24
4.3 Twitter sentiment - emotion variables based on Staiano & Guerini (2014) 25
4.4 Hourly log-differenced emotion variables based on Drummond (2004) 26
4.5 Log-returns of JY 29
4.6 Log-returns of S&P 30
4.7 Log-returns of NG 31
4.8 Log-returns of SV 32
4.9 Log-returns of CL 33
4.10 Log-returns of CD 34

5.1 Stability of the AR(1) model for JY returns, where the explanatory variable is the USD tag 39
5.2 Stability of the AR(1) model for JY returns, where the exogenous explanatory variable is negative mood (weekly logarithmic differenced values). 41
5.3 Plot of the predicted conditional volatility using one hour-lagged values of the negative mood and realized volatility of JPY. . . . 45
5.4 Volatility - realized and predicted conditional based on the model above for SV 47
5.5 Volatility - realized and predicted conditional based on the model with explanatory variable anger (three hours lagged daily logarithmic difference) for NG 48

6.1 Wavelet coherence of JY in pairs with the same Twitter sentiment as used in the table 5.1 58
6.2 Wavelet coherence of S&P log-returns (listed as the second variable) in pairs with emotions based on Drummond (2004) 60
6.3 Wavelet coherence of S&P close price (listed as the second variable) in pairs with emotions based Staiano & Guerini (2014) . . 61
6.4 Wavelet coherence of S&P log-returns (listed as the second variable) in pairs with moods based Drummond (2004), ... 62

Acronyms

BIC Bayesian information criterion

CD Canadian Dollar

CL Light Crude NYMEX

GCA Granger Causality Analysis

JY Japanese Yen

LL Log-likelihood

NG Natural Gas NYMEX

SV Silver COMEX

VAR Vector Autoregressive model

Master's Thesis Proposal

Author	Bc. Marek Fremunt
Supervisor	PhDr. Jozef Baruník, Ph.D.
Proposed topic	Predictability of security returns using Twitter sentiment

Topic characteristics The Efficient market hypothesis proposed by Fama (1970) assumes that investors value the assets in a rational way, while behavioral finance economists oppose this idea. Behavioral finance explores the human behavior and mood claiming that investors are not perfect homo economicus. Anything affecting human emotions and mood can make systematic errors in investors' decisions. This thesis aims to examine the relationship between Twitter sentiment and security returns as well as the capability to predict the volatility of financial instrument returns. Twitter sentiment is used as a good representative of investors' mood.

There have been several papers examining the influence of the internet and social media on stock returns. Some recent research show that sentiment tracking of social media may bring very promising results for stock returns predictions, for example Si et al. (2013), Rao and Srivastava (2012), Karabulut (2011) or Bollen et al. (2011). Drimpfl and Jank (2011) showed that Google search queries from Google Trends contain additional information about market volatility, which help to improve volatility forecasts in-sample and out-of-sample. Also measures of positive and negative media sentiment have significant relationships with stock returns as documented in Ferguson et al. (2011).

In my thesis we will analyze the content of 30 English written Tweets every second and collect the data hourly. In detail, we will measure ten basic emotions (happiness, caring, depression, inadequateness, fear, confusion, hurt, anger, loneliness, remorse) based on Vocabulary of emotion proposed by Drummond (2004). Moreover, we will measure positive and negative mood; positive, neutral and negative tone words and keywords related to specific securities and FOREX symbols (e.g. USD, GBPUSD, Microsoft or Nasdaq). The measurement of the keywords should work on the similar principle as Google search queries. The substantial part of my work will be modeling and forecasting three market variables - returns, volatility and trading volume with the above described Twitter data.

Hypotheses 1. Hypothesis #1: Twitter sentiment has no impact on security returns and its volatility (i.e. changes in mood do not affect investors' decision making processes in financial markets).
2. Hypothesis #2: Twitter keywords related to examined securities have no predictability power on returns or volatility.
3. Hypothesis #3: Hourly Twitter sentiment is capable to better predict the future returns in comparison to daily data

Methodology In my thesis we will use the Twitter data described above to predict the market returns, volatility and trading volume. For predicting securities' returns and trading volumes we will use linear models, for instance ARIMA models. When fitting ARIMA models, we will use Box-Jenkins methodology. We will also try to improve results by introducing a technique based on Wavelet transforms and ARIMA models. In case of the presence of nonlinearities in the data, we will introduce some nonlinear technique such as non-linear Support vector machine, Neural Network with wavelet decomposition or other relevant model.

The volatility of securities' returns will be predicted by the GARCH-type models. The GARCH family models are able to capture non-linear dynamics in the real life data, especially in financial markets. We will discuss and analyze both in-sample and out-of-sample forecast quality of the models.

Forecast evaluation of the models will be based on techniques like root square mean error, mean percentage error or mean absolute percentage error. Further, we will test the stability of selected models over time.

The first two hypotheses will be tested based on the outputs from my empirical analysis and Granger causality analysis. For the third hypothesis we will introduce analysis of daily data and compare the predictive power of both daily and hourly data based models.

Expected Contribution: In my thesis we will conduct a prediction of security returns and volatility. In contrast to other papers exploring the correlation between Twitter data and financial instrument returns, we will examine the impact of ten basic emotions using hourly data and a robust dataset consisting of text analysis from 108,000 Tweets collected per hour. Moreover, we will take into account frequencies of particular keywords related to specific securities and FOREX symbols, which should produce predictive and causative relationships with the returns. The Twitter data will be collected using my application programmed specially for this work.

This thesis will help to understand investors' sentiment, their behavior and will try to detect particular emotions affecting investors in their decision making processes. We would also like to explore the transmission speed of Twitter sentiment to security returns. Further, we will try to predict the volatility based on the collected Twitter data. Predicting volatility is very important in financial markets, since risk is usually expressed by volatility. The ability to predict volatility is useful, for example, in risk management, asset pricing, hedging etc.

Outline

1. Introduction
2. Literature review
3. Efficient market hypothesis and behavioral finance
4. Methodology
5. Data description
6. Empirical results
7. Conclusion and discussion of the results

Core bibliography

1. Barberis, N. & Thaler, R. (2003). "A survey of behavioral finance." *Handbook of the Economics of Finance* **1**: 1053-1128.

2. Bollen, J., Mao, H.,& Zeng, X. (2011). "Twitter mood predicts the stock market." *Journal of Computational Science* **2(1)**: 1-8.

3. Dimpfl, T., & Jank, S. (2011). "Can internet search queries help to predict stock market volatility? (No. 11-15)." *CFR working paper.*

4. Drummond T., (2004). "Vocabulary of Emotions [Online]." North Seattle Community College, [last viewed 7/2014]. Available from http://www.sba.pdx.edu/faculty/mblake/448/FeelingsList.pdf

5. Engle, R. F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation." *Econometrica: Journal of the Econometric Society*, 987-1007.

6. Fama, E. F. (1970). "Efficient capital markets: A review of theory and empirical work." *The journal of Finance* **25(2)**: 383-417.

7. Ferguson, N., Guo, J., Lam, H., & Philip, D. (2011). "Media sentiment and UK stock returns." *Working Paper.* Durham University, Durham.

8. Karabulut, Y. (2011). "Can Facebook predict stock market activity?." *SSRN eLibrary.*

9. Rao, T. & Srivastava, S. (2012). "Analyzing stock market movements using twitter sentiment analysis." In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2012) (pp. 119-123). IEEE Computer Society.

10. Shiller, R. J. (2003). "From efficient markets theory to behavioral finance." *Journal of economic perspectives*, 83-104.

11. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. & Deng, X. (2013). "Exploiting Topic based Twitter Sentiment for Stock Prediction." *In ACL* **(2)**:(pp. 24-29).

Author Supervisor

Chapter 1

Introduction

> *"A wonderful fact to reflect upon, that every human creature is constituted to be that profound secret and mystery to every other. A solemn consideration, when I enter a great city by night, that every one of those darkly clustered houses encloses its own secret; that every room in every one of them encloses its own secret; that every beating heart in the hundreds of thousands of breasts there, is, in some of its imaginings, a secret to the heart nearest it!"*
> Charles Dickens, A Tale of Two Cities

The Efficient market hypothesis proposed by Fama (1970) assumes that investors value the assets in a rational way, while behavioral finance economists oppose this idea. Behavioral finance explores the human behavior and mood claiming that investors are not perfect homo economicus. Anything affecting human emotions and mood can make systematic errors in investors' decisions. This thesis is located on the side of behavioral finance. We aim to examine the relationship between Twitter sentiment and financial instruments' returns as well as the capability to predict the volatility of securities' returns. We use Twitter sentiment as a good representative of investors' mood.

Psychologists suggest that people are either irrational or not fully rational beings. A theory of cognitive biases was introduced by a laureate of the Nobel Memorial Prize in Economic Sciences Daniel Kahneman, who focuses on psychology of judgment and decision-making, as well as behavioral economics. Cognitive biases are defined as tendencies to think in certain ways that can lead to systematic deviations from rationality. There have been more than 50 cognitive biases identified, which negatively influence human decision-making. Magical thinking presents a theory also showing humans' very limited rationality as shown in Zusne & Jones (2014). People tend to have irrational beliefs and to find causal relationships between two completely unrelated events. Magical thinking is related to believing in the intuition more than in a rational analysis. People are emotional beings, and emotions influence or often determine their behavior. Neuroscience came up with the somatic marker hypothesis, which underlines the crucial role of emotions in the ability to make fast and rational decisions in complex and uncertain situations. Bechara & Damasio (2005) show that rational decision-making deeply depends on prior accurate emotional processing. Through the somatic marker hypothesis they identify the influence of emotions and feelings, which can occur consciously or non-consciously.

Another important thing related to rationality is free will - do people have it? Psychologists in the last decades have proven that we do not fully have free will. Soon *et al.* (2008) described that we may often think we freely decide to do something, but in fact, we were determined to do that many seconds before we realized it.

Emotional contagion refers to the tendency for two individuals or social groups to emotionally converge. This term is very important for our thesis especially because of social networks. It implies some level on homogeneity of emotions within a social group. There have been many papers about this topic, for example Kramer *et al.* (2014) show that emotions expressed by others on Facebook influence our own emotions.

Based on arguments mentioned above, we cannot expect investors to behave and act in a fully rational way. They may, therefore, make systematic errors in their decision making processes and be influenced by emotions and moods. With the increasing role of the Internet and people's connectivity, as well as their activity on the social networks, the ability to track Internet sentiment becomes a very powerful and efficient tool for us. We believe that investors' sentiment can be tracked through proxy variables, for which we have chosen social sentiment indicators based on Twitter data.

We aim to explore and understand how are investors influenced by emotions, moods and sentiment in their decision making processes. We examine the relationships between sentiment variables and returns and volatility of several FOREX and futures symbols. We have chosen hourly data, while the related literature focuses on daily and weekly data. Our approach with using higher frequency of the data is novel. Moreover, we take into account frequencies of particular keywords related to specific securities and FOREX symbols, which may also produce causative relationships with the returns and volatility. The

Twitter data are collected using our own application programmed specially for this work based on analysis of more than 210,000,000 English written Tweets over the period of three months. Predicting volatility is very important in financial markets, since risk is usually expressed by volatility. The ability to predict volatility is useful, for example, in risk management, asset pricing, hedging, etc.

The empirical part of the thesis uses a wide scale of methodologies. Firstly, we conduct a Granger causality analysis. We used conditional and realized volatility models to predict the volatility. For predicting returns we chose ARIMA and GARCH models. To explore the explanatory power of Twitter sentiment variables we used Wavelet coherence.

The thesis is structured as follows. In chapter two we introduce the concept of behavioral finance, on which this thesis stands. We describe the main approaches to sentiment analysis. Then we introduce a wide scale of research from different fields related to Twitter sentiment. Finally, we summarize others' research based on social media and financial market. In chapter three we describe all methodologies applied in this work. We start with a Granger causality analysis, ARIMA models, Conditional Heteroscedastic models, realized volatility model and wavelets. In the following chapter we present our data collection method as well as the sentiment tracking and processing approach. All variables are defined and described as well as their statistics. The empirical results are shown in chapter five. We comment on the results about the impact of Twitter sentiment on securities' returns and volatility. The sixth chapter examines the explanatory power of our Twitter sentiment variables using Wavelet coherence. We show their performance in the time-frequency domain.

Chapter 2

Literature review

In this chapter we briefly describe the sentiment analysis and its application in the financial markets. In theory, sentiment analysis refers to the natural language processing and computational study of people's moods, emotions, opinions and attitudes at a certain time. It tends to generalize sentiment of a certain entity or a social group. With the increasing number of internet users, the social networks, as well as micro-blogs, have become a very interesting subject of research. Therefore the automatic text categorization is a powerful tool in a wide scale of fields, including the financial markets.

There are two main approaches for text-based sentiment analysis. The first one is based on machine learning algorithms. Hemalatha *et al.* (2013) define machine learning as a field of artificial intelligence designed and developed in order to enable computers to evolve behaviors based on empirical data. Further, the machine learning approach can automatically learn and recognize complex patterns as well as make decisions based on empirical data. Machine learning is divided into supervised and unsupervised learning techniques. Medhat *et al.* (2014) define the supervised methods as those utilizing a large number of labeled training documents, while the unsupervised methods are not using these labeled training documents. Supervised learning approach has four subsections: decision tree classifiers, linear classifiers, rule-based classifiers and probabilistic classifiers. The most popular are linear classifiers, which include support vector machines and neural network, and probabilistic classifiers. Probabilistic classifiers are able to predict a probability distribution over a set of classes. Probabilistic classifiers comprise Naïve Bayesian classification, Bayesian network and a maximum entropy classifier.

The other approach for text-based sentiment analysis is lexicon-based, which assumes that semantic orientation of a text is an averaged sum of the semantic orientation of its words, phrases and symbols. In the figure 2.1, based on Medhat *et al.* (2014), we show the scheme of sentiment classification techniques.

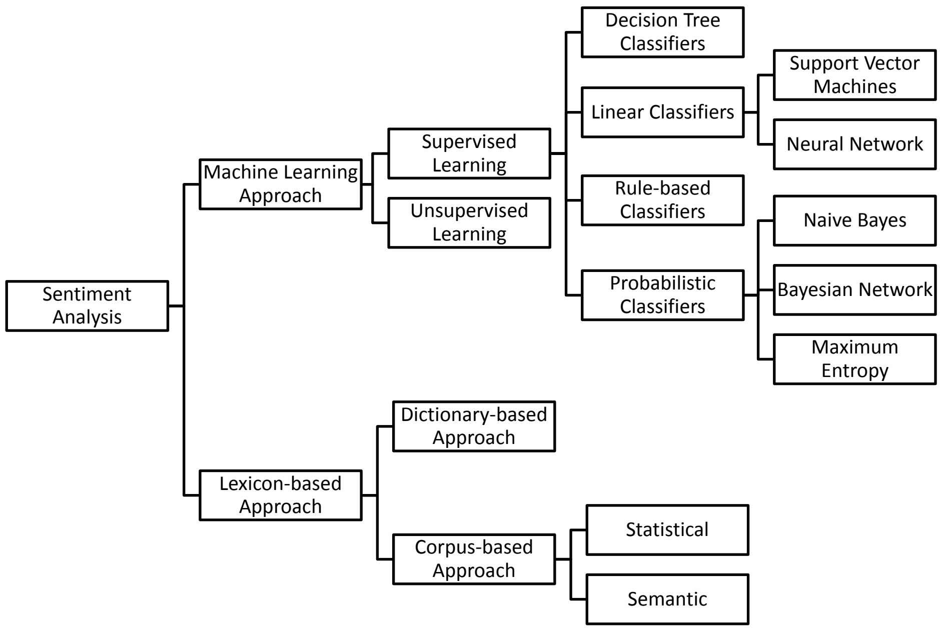


Figure 2.1: Sentiment analysis from Medhat *et al.* (2014)

Recently, with the increasing popularity of social networks and micro-blogs, sentiment analysis has become a popular and also an efficient tool for everyday life. Economists have not missed this new phenomenon either. One of the earliest attempts to use internet sentiment data for predicting stock prices was proposed by Tumarkin & Whitelaw (2001). They examined internet financial forums and were looking for a link between the content of postings and stock prices. Their results did not find a predictive power in message board activity on stock returns. Lately, Antweiler & Frank (2004) conducted similar research. They explored the effect of internet stock message boards on stock returns. The research was based on more than 1.5 million messages on two message boards for 45 companies from Dow Jones Industrial Average. They showed that there is a statistically significant effect but economically very small. Very promising results, in terms of predictive power, was brought about by Chen *et al.* (2014). They conducted a textual analysis of articles published on popular social media platforms for investors in the United States. They found that the information and opinions contained in both articles and commentaries strongly helped to predict future stock returns and earnings surprises.

A huge number of papers from various scale of fields explain the importance of microblogging platforms for opinion mining and sentiment analysis in today's world. Millions of users share their opinions and express their emotions and moods. The most popular microblogging platform is Twitter. The good thing on tracking the sentiment on Twitter is that its quick reaction to events. This real-time nature of Twitter was also important for the research conducted by Sakaki *et al.* (2010). They examined whether the activity of Twitter users was able to detect earthquake shakes. Twitter users were taken as sensors of an earthquake. Through an algorithm analyzing Tweets and a probabilistic spatiotemporal model, they are able to promptly detect an earthquake with seismic intensity scale 3 or more and its location. Twitter sentiment can be also used for predicting elections. Tumasjan *et al.* (2010) suggest on the case of German federal election that Twitter political sentiment may plausibly reflects the offline political landscape. They conducted a sentiment analysis on the sample of roughly 100,000 Tweets which containing a reference to either a political party or a politician.

First attempts to detect a causality between Twitter sentiment and securities' returns brought very promising results. Tayal & Komaragiri (2009) studied the influence of Twitter and blog sentiment on stocks' prices (Microsoft Corp. and Google Inc.) after Twitter's boom in early 2009. They measured the positivity and negativity of the postings. Their results claim that microblogs consistently outperformed blogs in their predictive power. Sprenger *et al.* (2014) examined roughly 250,000 stock-related Tweets on a daily basis in 2010. They used the Naïve Bayesian classification method to classify Tweets as either buy, hold or sell signals. They found that Twitter stock-related sentiment effects stock returns, and Tweets volume is able to predict next-day trading volume.

Bollen *et al.* (2011) investigated the effect of two mood tracking tools on the daily data of Dow Jones Industrial Average close price. The mood tracking tools were OpinionFinder, measuring positive and negative mood, and Google-Profile of Mood States, which measures six moods: alert, calm, happy, kind, sure and vital. In order to examine the effect of the mood tracking tools on changes in DJIA closing values, they used Granger causality analysis and a Self-Organizing Fuzzy Neural Network. They found that the predicting accuracy can be significantly improved by the inclusion of specific public mood. By the inclusion, they got a very high accuracy (87.6%) in predicting the daily up and down changes in the closing prices of DJIA.

Lately, research has confirmed the relation between Twitter sentiment and stock returns predictions. All this research works with lower frequency data, namely daily, weekly and monthly.

Si *et al.* (2013) collected 624,782 tweets over the period of 97 days via Twitter's REST API for streaming data, using symbols of the Standard & Poor's 100 stocks as keywords. Their Tweet-processing method consisted of topic based sentiment analysis using the Dirichlet Processes Mixture model. They showed that Twitter's topic based sentiment can improve the prediction accuracy.

Rao & Srivastava (2012) studied a period of 14 months, between June 2010 to July 2011, and collected roughly 4 million tweets. They focused on exploring the relationships between Twitter sentiment and volatility, trading volume as well as stock prices of DJIA, NASDAQ-100 and 13 other big cap technological stocks. For the Twitter classification, they used Naïve Bayesian classification determining tweets' value of positivity and negativity. Besides the total amount of tweets for an examined financial instrument, they introduce the variables for Bullishness and Agreement, both of which are computed just based on the level of positivity and negativity. Their results suggest strong dependencies between the market data and twitter sentiment.

Karabulut (2013) examines the influence of Facebook on the stock market returns. He found that Facebook's Gross National Happiness can help predict changes in daily returns as well as trading volume of the US stock market. Gross National Happiness is derived from positive and negative words in users' statuses.

There are also other relevant fields of research related to tracking the sentiment on the internet. News and media sentiment may be a very efficient tool, as shown e.g. in Chowdhury *et al.* (2014). Ferguson *et al.* (2011) tracked positive and negative UK media sentiment for UK companies and showed that they have significant relationships with the stock returns. Concretely, positive media sentiment in company-specific news articles positively influences company's stock returns and vice versa. According to their findings, a positive media sentiment is three times stronger than the negative on the day the news is published. On the following day, negative sentiment is almost twice as high as positive sentiment.

Google search queries from Google Trends may contain additional information about market volatility. Dimpfl & Jank (2012) prove that Google search queries can improve volatility forecasts in-sample and out-of-sample for different forecasting horizons (daily, weekly and monthly). They emphasize the ability of the search queries to predict volatility especially in high-volatility phases.

Chapter 3

Methodology

At the beginning of this chapter, we state three hypotheses exploring the impact of Twitter data on the financial markets. Then, we start with exploring the impact of Twitter sentiment variables on securities' returns and volatility. We examine their causative relationships on returns using Granger Causality Analysis (GCA). GCA is a very good tool for determining whether one time series can be used for forecasting another time series. It helps us to detect the significance and even predictive power of Twitter-based lagged variables. Based on this analysis, we can get a general notion about a basic relation between Twitter sentiment and market data. Further, we describe both linear and nonlinear models, for instance ARIMA and GARCH models. Securities' returns are modeled by ARIMA models with a Twitter sentiment variable as an exogenous explanatory variable. We also describe the stability of the models applying rolling analysis. Further, we present conditional volatility models. We add an exogenous explanatory variable into the variance equation of the GARCH model, whose performance is compared in the empirical results with the plain GARCH model. Moreover, we introduce realized volatility models, which deal with the downside of conditional volatility models using a proxy for the latent volatility by applying intraday high frequency data. We also add an exogenous explanatory variables, representing Twitter sentiment, into the HAR-RV model. Finally, we introduce wavelets and wavelet coherence. Wavelet coherence is considered as an efficient tool for explanatory power analysis that enables us to study the dependencies between two time series over time across different frequencies.

3.1 Hypotheses

In this theses we want to test following hypotheses:

1. Hypothesis #1: Twitter sentiment has no impact on security returns (i.e. changes in mood do not affect investors' decision making processes in financial markets).

2. Hypothesis #2: Twitter keywords related to examined securities have no predictability power on returns.

3. Hypothesis #3: Twitter sentiment has no impact on volatility of security returns.

3.2 Granger causality analysis

In order to explore the significance and predictive power of stationary lagged variables, we use the Granger causality test, proposed by Granger (1969). We test whether the Twitter sentiment variables Granger-cause the returns. The causality testing first involves running the bivariate Vector Autoregressive model (VAR) at an exact lag q with two endogenous variables, the security's returns (denoted y_{t-i}) and the Twitter sentiment variable (denoted x_{t-i}). In the next step we conduct an F-test on all lagged Twitter sentiment variables. The VAR model is defined as follows:

$$\begin{aligned} y_t &= \phi_1 + \textstyle\sum_{i=1}^{q} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i x_{t-i} + \epsilon_t \\ x_t &= \phi_2 + \textstyle\sum_{i=1}^{q} \gamma_i y_{t-i} + \sum_{i=1}^{q} \delta_i x_{t-i} + \epsilon_t \end{aligned} \tag{3.1}$$

The null hypothesis tests:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0 \tag{3.2}$$

Under the null hypothesis the restricted model has the form:

$$y_t = \phi + \sum_{i=1}^{q} \alpha_i y_{t-i} + \epsilon_t \tag{3.3}$$

Finally we compute the test statistics:

$$F = \frac{RSS_0 - RSS_1}{RSS_1/\left(T - q\right)} \tag{3.4}$$

where RSS_0 and RSS_1 stands for the residual sum of squares of restricted and unrestricted model respectively. T represents the length of the time series. Under the null $F \rightarrow \chi^2\left(q\right)$.

3.3 ARIMA models

ARIMA (p, d, q) are models widely used for linear modeling of securities' returns. They combine autoregressive and moving average models. The order p of autoregressive term can be written as:

$$r_t = \sum_{i=1}^{p} \rho_i r_{t-i} + \epsilon_t \tag{3.5}$$

The integration order term d represents an order of differencing and captures the stochastic trend. The order d is as low as possible to reach the stationarity. Moving average of the order q has the form:

$$r_t = \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3.6}$$

The ARIMA $(p, 0, q)$ model with exogenous explanatory variables has the form:

$$r_t = \sum_{i=1}^{p} \rho_i r_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{n} \alpha_i x_{i,t} + \epsilon_t \tag{3.7}$$

where $\epsilon_t \sim N\left(0, \sigma^2\right)$ and $x_{i,t}$ represents exogenous explanatory variables.

In this work we use ARIMA $(p, 0, q)$ models with one exogenous explanatory variable representing particular Twitter sentiment variable, denoted $x_{i,t}$:

$$r_t = \sum_{i=1}^{p} \rho_i r_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \alpha x_{i,t} + \epsilon_t \tag{3.8}$$

When fitting ARIMA, we use Box-Jenkins methodology (Box and Jenkins, 1976).

3.4 Conditional Heteroscedastic Models

Financial data has many important features, such as:

- volatility clustering - large and small price changes tend to occur in clusters,
- leverage effect - volatility tends to be higher after negative price shocks than after positive shocks of the same magnitude,
- leptokurtic distribution - log-returns are mostly heavy-tailed; in comparison to the Normal distribution there are more values around the sample mean and more extreme values, too.

In this section, we introduce models for volatility which aims to capture all of the above mentioned characteristics. The first basic model which tries to do that is:

$$r_t = \mu + \sigma_t \epsilon_t \tag{3.9}$$

Where σ_t is called the volatility process and represents a non-negative stochastic process, while $\{\epsilon_t\}$ is a sequence of i.i.d. random variables. It is also assumed that ϵ_t is i.i.d and $\epsilon_t \sim N(0, 1)$. Further, we assume that the time series σ_t and r_t are strictly stationary. We also assume $\mu = 0$, because we suppose that μ can be estimated.

Some literature also defines a new variable a_t called the mean corrected return, instead of assuming $\mu = 0$:

$$a_t = r_t - \mu_t \tag{3.10}$$

3.4.1 ARCH models

ARCH(1) model

The ARCH (Autoregressive Conditional Heteroskedasticity) model was introduced by Engle (1982). As shown above, a model for log-returns can have the form:

$$r_t = \sigma_t \epsilon_t \tag{3.11}$$

The first equation, often denoted as the mean equation, can have the form like the equation 3.11 or it can be, for example, written as simple AR(1) model:

$$r_t = \rho r_{t-1} + u_t \tag{3.12}$$

In our work we also use an exogenous explanatory variable in the mean equation, then the equation has the following form:

$$r_t = \sigma_t \epsilon_t + \gamma x_{i,t} \tag{3.13}$$

where $x_{i,t}$ stands for a concrete Twitter sentiment variable. The second equation, also called the variance equation, models the volatility process:

$$E\left(r_t\right)^2 = \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 \tag{3.14}$$

where α_i has to be non-negative for each $i > 0$.

Since one of the main purposes of this work is to reveal the influence of Twitter sentiment variables on volatility, we also use models with an exogenous explanatory variable in the variance equation.

$$E\left(r_t\right)^2 = \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \delta x_{i,t} \tag{3.15}$$

where $x_{i,t}$ denotes a particular Twitter sentiment variable.

If we add the assumption of normality, the model can be directly expressed in terms of information set available at time t: ψ_t, i.e. ϵ_t is conditionally normally distributed.

$$r_t | \psi_t \sim N\left(0, \sigma_t\right) \tag{3.16}$$

$$\sigma_t = \alpha_0 + \alpha_1 r_{t-1}^2 \tag{3.17}$$

Unconditional mean of r_t is zero, unconditional variance is $\frac{\alpha_0}{1-\alpha_1}$, process is stationary, if $\alpha_1 < 1$, the fourth moment is finite if $3\alpha_1^2 < 1$.

ARCH(q) model

The ARCH(q) model has q lagged terms in the equation 3.19. The first equation remains unchanged:

$$r_t = \sigma_t \epsilon_t + \gamma x_{i,t} \tag{3.18}$$

while the second one contains additional terms $\alpha_i r_{t-i}^2$:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + ... + \alpha_q r_{t-q}^2 + \delta x_{i,t} \tag{3.19}$$

where $x_{i,t}$ stands for a concrete Twitter sentiment variable and $\alpha_0 > 0$ and $\alpha_i > 0$ for all $i = 1, ..., q$. This system of two equations is then estimated by the maximum likelihood estimation.

3.4.2 GARCH models

The ARCH (q) model was further generalized by Bollerslev (1986) and introduced as the Garch model (General Autoregressive Conditional Heteroskedasticity), which is the most commonly used model in financial series. GARCH models solve one of the weaknesses of the ARCH model. The ARCH model assumes that positive and negative shocks have the same effect on volatility, while the empirical findings show rather the leverage effect. The GARCH models' conditional variance depend on both the magnitude and the sign of the log-returns.

GARCH(1,1)

GARCH(1,1) model is the most frequently used model from the GARCH family.

$$r_t = \sigma_t \epsilon_t + \gamma x_{i,t} \tag{3.20}$$

The second one contains, in comparison to ARCH(1) model, additional term $\beta_1 \sigma_{t-1}^2$:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \delta x_{i,t} \tag{3.21}$$

where $\alpha_0 > 0$, $\alpha_1 > 0$, $\beta_1 > 0$ and $(\alpha_1 + \beta_1) < 1$.

GARCH(p,q) model

The GARCH(p,q) model has the form:

$$r_t = \sigma_t \epsilon_t + \gamma x_{i,t} \tag{3.22}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + ... + \alpha_q r_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + ... + \beta_p \sigma_{t-p}^2 + \delta x_{i,t} \tag{3.23}$$

where $\alpha_0 > 0$, $\alpha_i > 0$ for all $i = 1, ..., q$ and $\beta_j > 0$ for all $j = 1, ..., p$. Also the following condition has to be fulfilled:

$$\sum_{i=1}^{max(p,q)} \alpha_i + \beta_i < 1 \tag{3.24}$$

If $q = 0$, then GARCH(p,q) model reduces to simple ARCH(p) model.

3.5 Realized Volatility

3.5.1 HAR models

Corsi (2004) introduces realized volatility model - Heterogeneous Autoregressive model. The downside of conditional variance models is that they have problems to replicate main empirical features of financial data and the estimations are non trivial. Corsi (2004) suggests an alternative approach. He constructs a proxy, denoted as Realized Volatility, for the latent volatility by using intraday high frequency data. This model was inspired by Müller *et al.* (1993) and their hypothesis of a heterogeneous market and the HARCH model proposed by Müller *et al.* (1997).

3.5.2 HAR-RV model

Following Corsi (2004) we introduce the HAR-RV model. Let $\widetilde{\sigma}_t^{(\cdot)}$ be the partial volatility generated by a certain market component. Further, we assume a hierarchical model with only three volatility components. They correspond to time horizons of one day $\widetilde{\sigma}_t^{(d)}$, one week $\widetilde{\sigma}_t^{(w)}$ an one month $\widetilde{\sigma}_t^{(m)}$.

We assume just three components because of simplification of a hierarchical process where at each level of time scale the future partial volatility depends on the past volatility with the same time scale as well as on the partial volatility at the next higher level of the cascade.

Moreover, we assume that the high frequency process is determined by the highest volatility component in the cascade with $\widetilde{\sigma}_t^{(d)} = \sigma_t^{(d)}$ the daily integrated volatility. Then we get the process:

$$r_t = \sigma_t^{(d)} \epsilon_t \tag{3.25}$$

where ϵ_t is independently normally distributed with zero mean and unit variance.

In order to model the unobserved partial volatility processes $\widetilde{\sigma}_t^{(d)}$ at each level of the cascade, it is assumed to be a function of the past realized volatility experienced at the same time scale and of the expectation of the next period values of the longer term partial volatilities. Then we get the cascade model of three equations:

$$\begin{aligned}
\widetilde{\sigma}^{(m)}_{t+1m} &= c^{(m)} + \phi^{(m)} + RV^{(m)}_t + \widetilde{\omega}^{(m)}_{t+1m} \\
\widetilde{\sigma}^{(w)}_{t+1w} &= c^{(w)} + \phi^{(w)} + RV^{(w)}_t + \gamma^{(w)} E_t\left[\widetilde{\sigma}^{(m)}_{t+1m}\right] + \widetilde{\omega}^{(w)}_{t+1w} \\
\widetilde{\sigma}^{(d)}_{t+1d} &= c^{(d)} + \phi^{(d)} + RV^{(d)}_t + \gamma^{(d)} E_t\left[\widetilde{\sigma}^{(w)}_{t+1w}\right] + \widetilde{\omega}^{(d)}_{t+1d}
\end{aligned} \tag{3.26}$$

where $RV^{(m)}_t$, $RV^{(m)}_t$ and $RV^{(m)}_t$ stand for daily, weekly and monthly observed realized volatility. The innovations terms $\omega^{(m)}_{t+1m}$, $\omega^{(w)}_{t+1w}$ and $\omega^{(d)}_{t+1d}$ are contemporaneously and serially independent. The ex-post daily, weekly and monthly realized volatilities are defined as follows:

$$\begin{aligned}
RV^{(d)}_t &= \textstyle\sum_{j=1}^{M} r^2_{t,j} \\
RV^{(w)}_t &= \tfrac{1}{5}\left(RV^{(d)}_t + RV^{(d)}_{t-1} + ... + RV^{(d)}_{t-4}\right) \\
RV^{(m)}_t &= \tfrac{1}{22}\left(RV^{(d)}_t + RV^{(d)}_{t-1} + ... + RV^{(d)}_{t-21}\right)
\end{aligned} \tag{3.27}$$

Where M represents the number of observed daily periods, which equals in our case to 24, since we work with hourly data. Note that we use the later definition of daily realized volatility from Corsi *et al.* (2008). Next, we can simplify the system of equations 3.26 into:

$$\sigma^{(d)}_{(t+1d)} = c + \beta_d RV^{(d)}_t + \beta_w RV^{(w)}_t + \beta_m RV^{(m)}_t + \widetilde{\omega}^{(d)}_{t+1d} \tag{3.28}$$

Further, we can write ex-post $\sigma^{(d)}_{(t+1d)}$ to be equal to the realized daily volatility plus daily volatility measurement error:

$$\sigma^{(d)}_{(t+1d)} = RV^{(d)}_{(t+1d)} + \omega^{(d)}_{t+1d} \tag{3.29}$$

By substituting 3.29 into 3.28, we finally get the HAR-RV model:

$$RV^{(d)}_{(t+1d)} = c + \beta_d RV^{(d)}_t + \beta_w RV^{(w)}_t + \beta_m RV^{(m)}_t + \widetilde{\omega}^{(d)}_{t+1d} \tag{3.30}$$

In our thesis we want to use the HAR-RV model with exogenous explanatory variables representing Twitter sentiment. The model has the form:

$$\begin{aligned}
&RV^{(d)}_{(t+1d)} = c + \beta_d RV^{(d)}_t + \beta_w RV^{(w)}_t + \beta_m RV^{(m)}_t + \widetilde{\omega}^{(d)}_{t+1d} + \\
&+\gamma_1 ERV^{(d)}_{t-1} + \gamma_2 ERV^{(w)}_{t-1}\epsilon_{1,t} + \gamma_3 ERV^{(m)}_{t-1}
\end{aligned} \tag{3.31}$$

where variables $ERV^{(d)}_{t-1}$, $ERV^{(w)}_{t-1}$ and $ERV^{(m)}_{t-1}$ stand for daily, weekly and monthly realized volatility of the hourly log-differenced values of a particular Twitter sentiment variable. For clarity, the daily, weekly and monthly realized volatilities of a particular Twitter sentiment variable are defined as follows:

$$\begin{aligned} ERV_t^{(d)} &= \textstyle\sum_{j=1}^{24} s_{t,j}^2 \\ ERV_t^{(w)} &= \tfrac{1}{7}\left(ERV_t^{(d)} + ERV_{t-1}^{(d)} + ... + ERV_{t-6}^{(d)}\right) \\ ERV_t^{(m)} &= \tfrac{1}{28}\left(ERV_t^{(d)} + ERV_{t-1}^{(d)} + ... + ERV_{t-27}^{(d)}\right) \end{aligned} \tag{3.32}$$

where variable $s_{i,t}$ represents log-differenced values of a particular Twitter sentiment variable (i). $s_{i,t}$ is calculated:

$$s_{i,t} = log\left(Value_{i,t}\right) - log\left(Value_{i,t-1}\right) \tag{3.33}$$

$s_{i,t}$ is the hourly log-differenced value of the i-th Twitter sentiment variable and $Value_{i,t}$ stands for the value of the variable.

3.6 Wavelets

In this section we introduce a wavelet and wavelet coherence, which we use to detect a causality between Twitter sentiment and financial market data. The main difference between Fourier and wavelet analysis is that Fourier, composed from sinusoids of different frequencies, does not have a limited duration. Fourier analysis transforms the signal from time-based domain into frequency domain and the time information is lost, while the wavelet transform is a time-frequency analysis of a signal. A wavelet $\psi \in L^2(\mathbb{R})$ is a function defined as:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \tag{3.34}$$

The wavelet is normalized $\|\psi\|^2 = 1$ and has zero mean. The wavelet function is localized in the time domain via translations of the mother wavelet and in the frequency domain by dilating the wavelet $\psi(t)$. The scale parameter s determines how the wavelet is dilated while the location in the time domain is determined by location (time shift) parameter u.

3.6.1 Continuous Wavelet Transform

Continuous Wavelet Transform has the form:

$$W_x(u,s) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \overline{\psi\left(\frac{t-u}{s}\right)} dt \tag{3.35}$$

where $x(t) \in L^2(\mathbb{R})$ is the examined time series, ψ a specific wavelet, u and s are time shift and scale respectively. Hence the term $\psi\left(\frac{t-u}{s}\right)$ represents a shifted and scaled version of a mother wavelet $\psi(t)$. In order to be able to reconstruct a time series from its wavelet transform, the wavelet admissibility condition has to be fulfilled:

$$C_\psi = 2\pi \int_{-\infty}^{\infty} \frac{\left|\widehat{\psi}(f)\right|^2}{f} df < \infty \tag{3.36}$$

where $\widehat{\psi}(f)$ is the Fourier transformation of ψ. The Fourier transform of the wavelet ψ can be written as:

$$\widehat{\psi}(f) = \int_{-\infty}^{\infty} \psi(t) e^{-i2\pi f t} dt \tag{3.37}$$

The wavelet admissibility condition implies that a function $\psi \in L^2(\mathbb{R})$ has zero mean on $(-\infty, \infty)$, i.e.:

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0 \tag{3.38}$$

There are complex wavelets, which have both real and imaginary parts. Complex wavelets enable us to separate the amplitude and phase components within the signal. In our work we use complex Morlet wavelet, defined e.g. in Addison (2002) as:

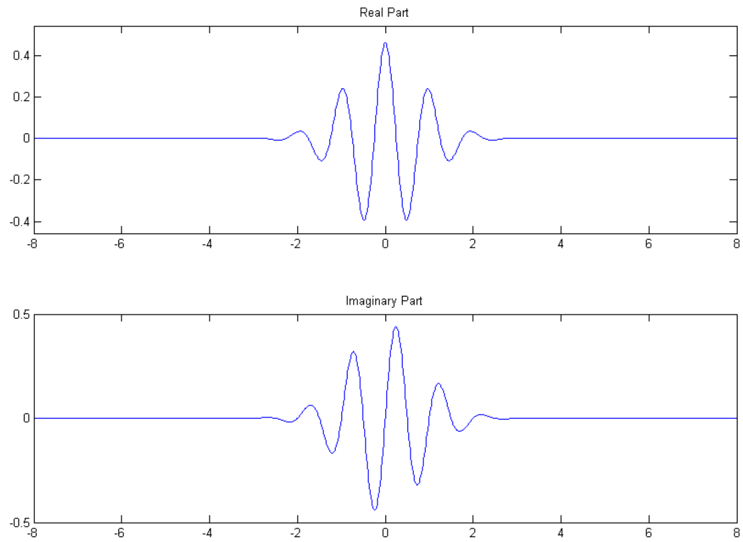
$$\psi(t) = \frac{1}{\sqrt[4]{\pi}} \left(e^{i2\pi f_0 t} - e^{-\frac{(2\pi f_0)^2}{2}}\right) e^{-\frac{t^2}{2}} \tag{3.39}$$

where parameter f_0 denotes the central frequency of the wavelet. The term $e^{-\frac{(2\pi f_0)^2}{2}}$ is a correction term, correcting the non-zero mean of the complex sinusoid. This term becomes negligible if we set $f_0 = 6$, as often used in literature. Hence we can write the mother wavelet as:

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}} e^{i2\pi f_0 t} e^{-\frac{t^2}{2}} \tag{3.40}$$

In the figure 3.43 we plot both real and imaginary parts of the complex Morlet wavelet with the central frequency parameter set to 6.

Figure 3.1: Plot of the Complex Morlet Wavelet with $f_0 = 6$



Source: author's computations.

3.6.2 Wavelet Coherence

Following Torrence & Webster (1999) and Grinsted *et al.* (2004) we introduce the wavelet coherence of two time series. Let $x(t)$ and $y(t)$ be time series with wavelet transforms $W_x(u,s)$ and $W_y(u,s)$, where s is the scale and n is the time index. Then we can define the cross wavelet spectrum as:

$$W_{xy}(u,s) = W_x(u,s)\,W_y^*(u,s) \tag{3.41}$$

where * denotes (a) complex conjugate. The squared wavelet coherence of two time series is defined as:

$$R^2(u,s) = \frac{\left|S\left(s^{-1}W_{xy}(u,s)\right)\right|}{S\left(s^{-1}\left|W_x(u,s)\right|^2\right)S\left(s^{-1}\left|W_y(u,s)\right|^2\right)} \tag{3.42}$$

S is a smoothing operator, which can be written as:

$$S(W) = S_{scale}\left(S_{time}\left(W(u,s)\right)\right) \tag{3.43}$$

where S_{scale} stands for smoothing along the wavelet scale axis and S_{time} smoothing in time. The definition (see equation: 3.43) resembles a correlation coefficient, $0 \leq R^2(u,s) \leq 1$, and we can think of the wavelet coherence as a localized correlation in the time frequency space (Grinsted *et al.* (2004)).

3.6.3 Wavelet coherence phase

The wavelet-coherency phase is defined as:

$$\phi(u,s) = tan^{-1}\left(\frac{\Im\left\{S\left(s^{-1}W_{xy}(u,s)\right)\right\}}{\Re\left\{S\left(s^{-1}W_{xy}(s)\right)\right\}}\right) \tag{3.44}$$

For computing and plotting the wavelet coherence with phase we use software Matlab and package proposed by Grinsted *et al.* (2004). The arrows in the figure represent the relative phase between the two signals as a function of scale and position. A rightward (leftward) arrow indicates that the time series are in-phase (anti-phase) or positively correlated (anti-correlated). The downward and upward direction of arrows indicates whether the first time series lead by 90° the second or vice versa. Usually the vertical and horizontal directions of arrows are combined.

Chapter 4

Data description

4.1 Twitter Sentiment tracking methods

Twitter is a popular online social network where every user can post short 140-character messages called "tweets". Tweets, except the protected ones, are available to read for every internet user - not just for the Twitter users. According to Twitter statistics, over 500 million tweets are sent per day. There are 284 million monthly active users and 23% of them are from the USA.

For our work, we have obtained every second 25 English written Tweets. From these Tweets we extract the Twitter sentiment on an hourly basis. In this work, we have computed the Twitter Sentiment for the period between July 1, 2014 and October 9, 2014. The data set contains over 210 million English written Tweets.

This diploma thesis uses Twitter streaming API framework for Tweet collection. Twitter Streaming API returns a real-time random sample of all public Tweets. We can set several parameters, such as the language and locations (just geolocated Tweets falling within the requested bounding are received). We have set the language filter to English, hence we get only stream Tweets detected to be in the English language. The system of scripts and programs are written in PHP, C++ and Java. First, the system collects Tweets from Twitter streaming API and save them into a database. From the database the data are extracted and processed as described in the section below. The output is saved into MySQL database, from which the data are easily accessible for us, e.g. as a CSV file.

4.2 Twitter Sentiment text analysis

Firstly, we have preprocessed the collected Tweets by tokenization and part-of-speech tagging. We have used product from Gimpel *et al.* (2011), which provides a very fast and robust tool. Secondly, we conduct a lexical analysis of the preprocessed Tweets. We track specific words from several lexicons. The first one is a very comprehensive Lexicon called DepecheMood: a Lexicon for Emotion Analysis designed by Staiano & Guerini (2014). DepecheMood is available in three versions: raw frequencies, normalized frequencies and term frequency-inverse document frequency. We have used the normalized frequencies and term frequency-inverse document frequency. The Lexicon contains over 37,000 words; each word has scores for the following emotions: afraid, amused, angry, annoyed, don't care, happy, inspired and sad.

The next lexicon is the Vocabulary of emotions proposed by Drummond (2004) containing keywords for ten basic human emotions: happiness, caring, depression, inadequateness, fear, confusion, hurt, anger, loneliness, remorse. The words from this lexicon are divided into three groups - strong, medium and light emotions. Further, we detect positive, negative and neutral tones of words, positive and negative moods based on Scott (2011). The keywords related to specific securities and FOREX symbols (e.g. USD, GBPUSD, oil or Nasdaq) are also included in this work. The idea behind the measurement of the keywords is that it should work on the similar principle as Google search queries and give some explanatory power.

4.3 Twitter Sentiment Extraction and Processing

From each Tweet we have extracted and processed every word contained in all the lexicons. In the lexicon DepecheMood there are eight emotional coefficients assigned for each word. When processing the words from Tweets, we sum all the coefficients from used words for each emotion separately. We collect the data on an hourly basis, so in final we get eight values of emotions every hour. When processing Tweets for the Vocabulary of emotions, we sum up the coefficient of followed words, where each group of emotions carry different weight. Words in the group strong emotions have coefficient 1, medium emotions 0.7 and light emotions 0.3. For the hourly values for the positive, negative and neutral tones of words, positive and negative moods are simply summed up, when each word has the same weight 1. The word extracting algorithm is able to recognize both the plural form of the nouns and all forms of the verbs, i.e. the third person singular, -ing form, past tense and past participle.

4.4 Twitter Sentiment variables

In figures 4.3 - 4.3 we show the time series plot of our collected raw data. From the figures we can see that the Twitter sentiment variables are all seasonal and non-stationary. To avoid problems with working with non-stationary time series, e.g. a spurious regression, we process all the variables using hourly, daily and weekly log-differences as shown in the equation 4.1. In the table A.3 we present the results of tests related to stationarity, concretely Augmented Dickey-Fuller test and KPSS test. The tests' results suggest that Twitter sentiment raw data are not stationary, while log-differences of these variables are stationary. In the table below, we present all raw Twitter-based variables. Note that in this thesis we only use their hourly, daily and weekly log-differenced values on an hourly frequency. The formula is given by the equation 4.1.

Table 4.1: Twitter sentiment variables

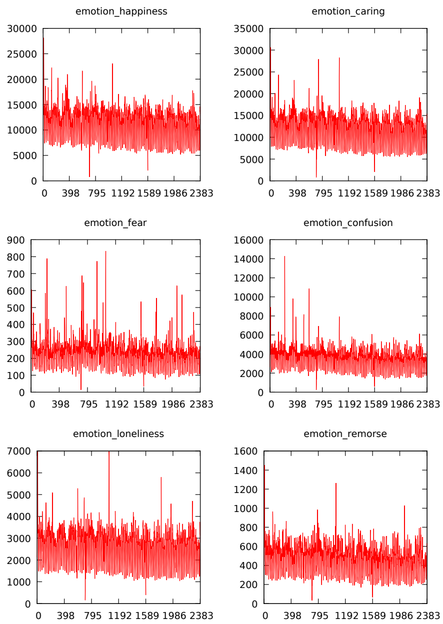
Emotions I Drummond (2004)	Emotions II Staiano & Guerini (2014)	Moods and Tones Drummond (2004)	Tags
happiness	afraid	mood positive	USD
caring	amused	mood negative	NASDAQ
depression	angry	tone positive	silver
inadequateness	annoyed	tone neutral	oil
fear	don't care	tone negative	gas
confusion	happy		CAD
hurt	inspired		USDCAD
anger	sad		
loneliness			
remorse			

All the variables are processed by the following formula:

$$s_{i,t}^{j} = log\left(Value_{i,t}\right) - log\left(Value_{i,t-j}\right) \tag{4.1}$$

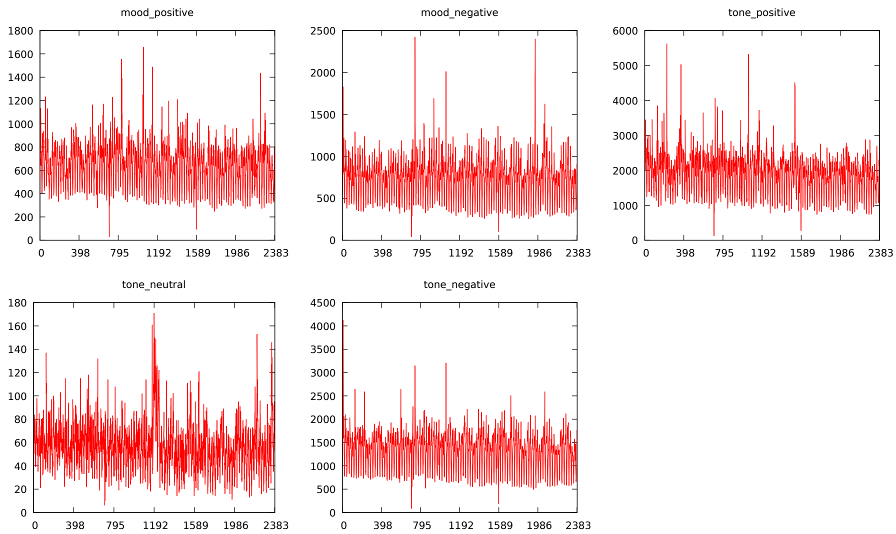
where $s_{i,t}$ is the j-hour logarithmic differenced value of the i-th Twitter sentiment variable and $Value_{i,t}$ stands for the value of the variable. As we have already mentioned, j equals to 1, 24 and 168.

Figure 4.1: Twitter sentiment - emotion variables based on Scott (2011)



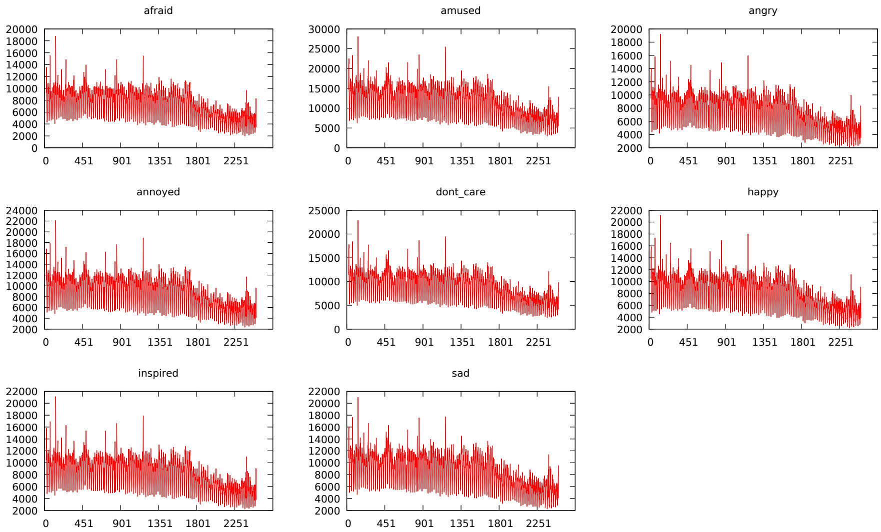
Source: author's computations.

Figure 4.2: Twitter sentiment - mood and tone variables based on Scott (2011)



Source: author's computations.

Figure 4.3: Twitter sentiment - emotion variables based on Staiano & Guerini (2014)



Source: author's computations.

4.5 Adjusting the time-frequencies of the Twitter sentiment and financial market data

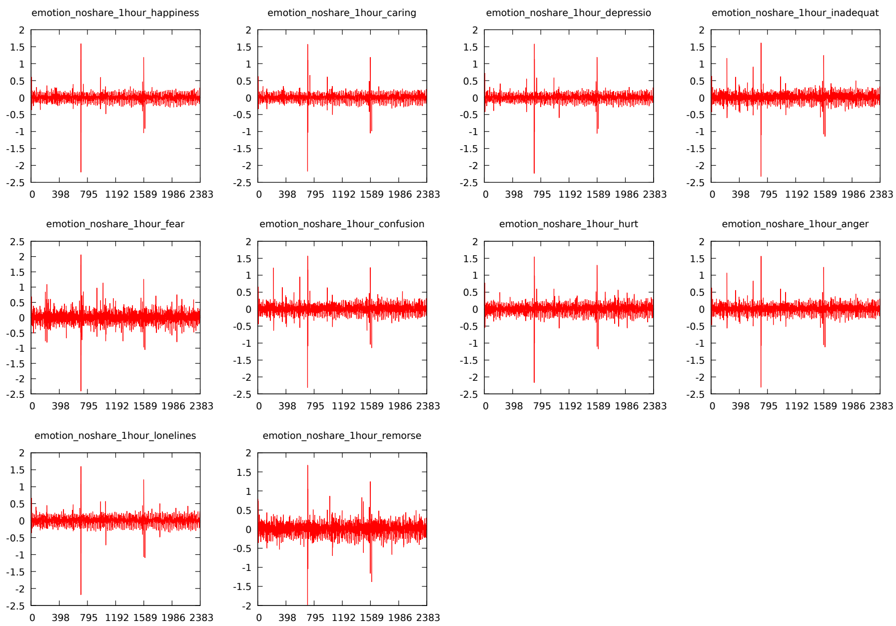
Since the securities we are using are not traded non-stop, we have to deal with a different length of Twitter sentiment and financial market data. We solve this problem by applying a special formula, based on weighted average, for the first hour of Twitter sentiment variables after a break in financial data. So that they contain information about the time when the trading was stopped, which should reveal in the prices after market opening.

4.6 Statistics of the Twitter variables

In the figure 4.4 below, we illustrate the performance of hourly log-differenced emotional variables. On the graphs we can see that through differencing we got rid of seasonality in the data. As the results of Augmented Dickey-Fuller test and KPSS test presented in the table A.4 suggest, the log-differenced variables are stationary.

We summarize the statistics of the positive, negative and neutral tones of words, positive and negative moods based on Scott (2011) in the table 4.2. The average values of logarithmic differences of mood and tones are close to zero, as expected. We can see the variable, among moods and tones, with the highest

Figure 4.4: Hourly log-differenced emotion variables based on Drummond (2004)



Source: author's computations.

value of the standard deviation. The variable is weekly logarithmic difference of the mood positive. It suggests that there are high fluctuations over time. The daily logarithmic differences are all leptokurtic, most of hourly and daily differenced variables have very high values of the kurtosis, whereas the weekly differenced perform with the lower values.

Table 4.2: Statistics of the mood and tone variables

Variable	Mean	Min	Max	Std. Dev.	Skewness	Kurtosis
mood positive	656.91	92	1659	178.47	0.11827	0.75339
mood negative	757.04	102	2423	253.36	0.70114	3.3727
tone positive	1936.1	274	5621	556.57	0.60562	3.0853
tone neutral	53.928	11	132	18.711	0.5382	0.76894
tone negative	1364.1	181	4121	408.79	0.26477	1.9084
mood positive (1H)	0.00104	-2.5543	2.7815	0.1803	1.1642	47.8474
mood negative (1H)	0.00027	-2.1300	1.7120	0.1777	-0.0576	17.6102
tone positive (1H)	0.00046	-2.1196	1.5966	0.1661	-0.0913	19.8877
tone neutral (1H)	0.00069	-1.8788	1.8506	0.2492	0.0243	4.7264
tone negative (1H)	-0.00002	-2.1379	1.7279	0.1628	-0.1529	23.7492
mood positive (1D)	0.02177	-2.6192	3.1489	0.3354	4.8030	40.1662
mood negative (1D)	0.00515	-2.1891	2.8129	0.2150	1.1881	26.6620
tone positive (1D)	0.00756	-2.5232	2.6199	0.2306	1.1785	20.8058
tone neutral (1D)	0.01594	-1.6964	2.4963	0.3526	1.2125	7.1394
tone negative (1D)	-0.00234	-2.3346	2.6459	0.1884	0.9580	40.6696
mood positive (1W)	0.16699	-3.3565	4.3847	0.6892	2.7040	7.8295
mood negative (1W)	0.04655	-3.0139	2.7093	0.3220	0.8002	10.9864
tone positive (1W)	0.06037	-2.8746	2.9462	0.3555	1.2807	8.0726
tone neutral (1W)	0.11965	-2.3168	3.1355	0.5540	1.6540	4.2636
tone negative (1W)	-0.00683	-2.9842	2.5987	0.2616	0.1526	18.7585

***(1H) means hourly log-differences, (1D) means daily log-differences, (1W) means weekly log-differences*

Source: author's computations.

We summarize the statistics of the emotion variables based on Staiano &

Guerini (2014) in table 4.3. The data suggests that the most frequent emotion on Twitter is amused and the least frequent is afraid. From the changes in emotions we can see that the most changeable emotion (with the highest value of the standard deviation) is afraid. The average values of the logarithmic differences are very close to zero, as would be expected. The daily logarithmic differences are all leptokurtic, which corresponds with the kurtosis of most of the financial data. The statistics of the other variables are shown in the appendix, in the tables A.1 and A.2.

Table 4.3: Statistics of the emotion variables based on Staiano & Guerini (2014)

Variable	Mean	Min	Max	Std. Dev.	Skewness	Kurtosis
AFRAID	7230.39	1998.12	18901.74	2403.24	0.1142	-0.5012
AMUSED	11442.22	3021.20	28456.75	3897.23	0.0974	-0.6166
ANGRY	7460.36	2084.16	19550.55	2460.23	0.0968	-0.4908
ANNOYED	8692.65	2397.78	22635.22	2885.28	0.0971	-0.5067
DON'T CARE	8873.51	2419.79	23400.01	2994.33	0.1332	-0.4306
HAPPY	8142.41	2236.82	21592.85	2727.01	0.1212	-0.4375
INSPIRED	8106.38	2245.49	21594.53	2701.59	0.1146	-0.4387
SAD	8431.06	2247.64	21281.35	2874.78	0.0918	-0.6535
AFRAID (1H)	0.0000051	-0.0611	0.0688	0.0147	0.0743	2.4350
AMUSED (1H)	0.0000014	-0.0689	0.0774	0.0125	0.3328	4.6961
ANGRY (1H)	0.0000020	-0.0542	0.0465	0.0079	-0.4067	5.5538
ANNOYED (1H)	-0.0000001	-0.0237	0.0253	0.0050	-0.1960	2.0445
DON'T CARE (1H)	-0.0000051	-0.0271	0.0343	0.0061	0.1118	2.4922
HAPPY (1H)	-0.0000034	-0.0395	0.0348	0.0062	-0.2161	4.1496
INSPIRED (1H)	-0.0000054	-0.0312	0.0282	0.0063	-0.1026	1.5948
SAD (1H)	0.0000057	-0.0615	0.0697	0.0105	0.3627	6.0301
AFRAID (1D)	-0.0062637	-0.9334	0.7310	0.1421	-0.2032	5.6174
AMUSED (1D)	-0.0066825	-0.9231	0.7590	0.1441	-0.2210	5.7526
ANGRY (1D)	-0.0064094	-0.9094	0.7362	0.1394	-0.1938	5.8717
ANNOYED (1D)	-0.0064676	-0.9112	0.7624	0.1398	-0.2080	6.0692
DON'T CARE (1D)	-0.0066815	-0.9158	0.7755	0.1422	-0.1809	5.9376
HAPPY (1D)	-0.0065578	-0.9150	0.7408	0.1414	-0.2131	5.9031
INSPIRED (1D)	-0.0065531	-0.9011	0.7617	0.1403	-0.2064	6.0495
SAD (1D)	-0.0066221	-0.9253	0.7174	0.1436	-0.2512	5.7351
AFRAID (1W)	0.0013639	-0.1064	0.0966	0.0244	-0.1549	0.6865
AMUSED (1W)	-0.0009939	-0.0940	0.0957	0.0180	0.1153	2.4877
ANGRY (1W)	0.0007327	-0.0680	0.0529	0.0119	-0.3741	2.7475
ANNOYED (1W)	0.0006201	-0.0243	0.0282	0.0068	-0.0627	0.7796
DON'T CARE (1W)	-0.0005955	-0.0529	0.0505	0.0091	-0.0553	2.7737
HAPPY (1W)	0.0001148	-0.0443	0.0387	0.0084	0.0002	1.5727
INSPIRED (1W)	0.0001622	-0.0375	0.0364	0.0097	-0.0081	0.6559
SAD (1W)	-0.0007644	-0.0657	0.0651	0.0145	0.0819	2.0975

(1H) means hourly log-differences, (1D) means daily log-differences, (1W) means weekly log-differences

Source: author's computations.

4.7 Financial Market Data

As we have already mentioned, we use the hourly data. The reason for choosing hourly data is that we aim to bypass problems with microstructure noise, which would be present in the case of using higher frequency. Microstructure noise can make high frequency estimates very unstable, thus the analysis would be biased. We also believe that hourly data are optimal to combine with Twitter sentiment data. Lower frequencies would be also possible to use with Twitter sentiment data, as shown in the literature review, but we aim to utilize the data on a high frequency data, hourly data are convenient for us. In this thesis we work with a various scale of securities, FOREX - CME Exchange Canadian Dollar and Japanese Yen, Metal Futures - Silver COMEX), Energy Futures (Light Crude NYMEX and Natural Gas NYMEX), and S&P 500 Cash Index.

For all securities we compute the returns r_t, which are defined:

$$r_t = log\left(Close_t\right) - log\left(Close_{t-1}\right) \tag{4.2}$$

computed as a logarithmic difference between current closing value $Close_t$ and one-hour delayed $Close_{t-1}$.

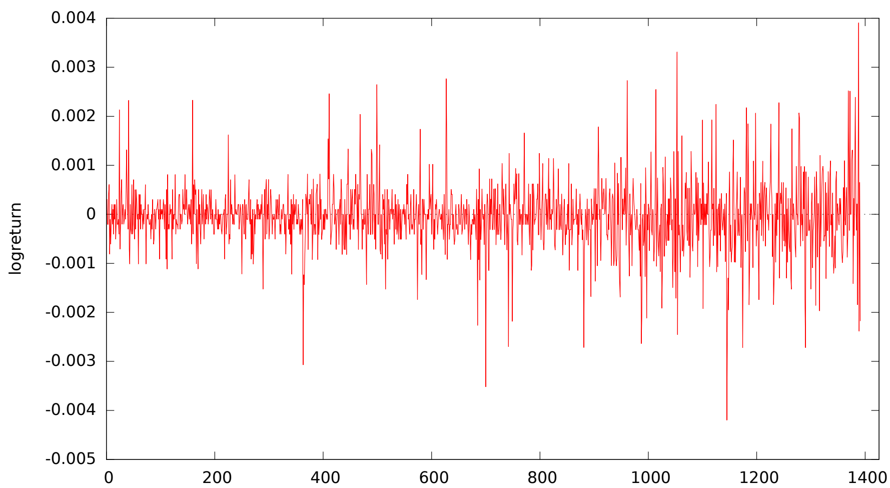
The results of Augmented Dickey-Fuller test and KPSS test presented in the table A.5 suggest that log-returns of all examined securities are stationary.

4.8 Statistics of the securities' returns

4.8.1 JY

Japanese Yen is the third most traded currency. JY consists of hourly data from USD/JPY reference rate. In the figure below we present the evolution of JY log-returns over the examined period, which begins on July 1, 2014 and ends on October 9, 2014.

Figure 4.5: Log-returns of JY



Source: author's computations.

4.8.2 S&P

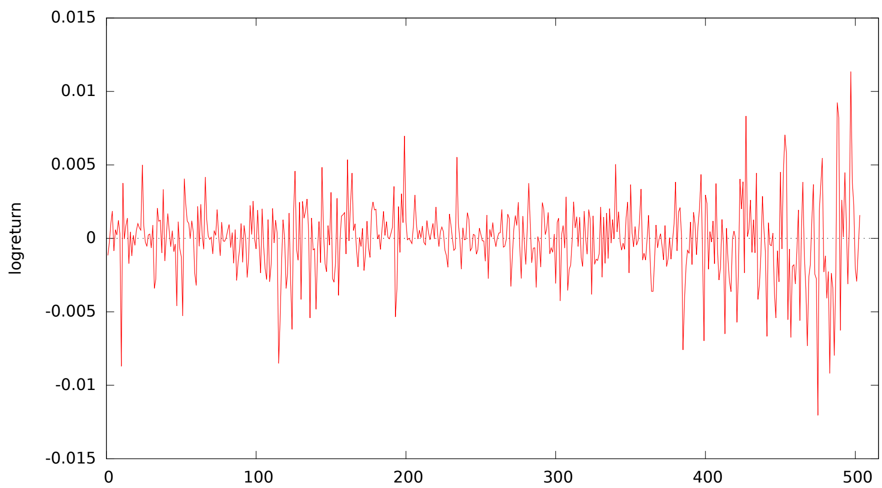
The Standard & Poor 500 index, denoted in this work as S&P, is composed by the 500 largest companies having common stock listed on the NASDAQ or NYSE. The index, founded in 1957, is considered as one of the best representations the U.S. stock market. In this work we use S&P 500 futures, traded on the Chicago Mercantile Exchange (CME).

Table 4.4: Statistics of JY

Mean	-4.7561e-005
Median	0.00000
Minimum	-0.0042000
Maximum	0.0039080
Standard deviation	0.00069435
C.V.	14.599
Skewness	0.025349
Ex. kurtosis	4.8230
5% percentile	-0.0010940
95% percentile	0.00097800
Interquartile range	0.00063200

Source: author's computations.

Figure 4.6: Log-returns of S&P



Source: author's computations.

Table 4.5: Statistics of S&P

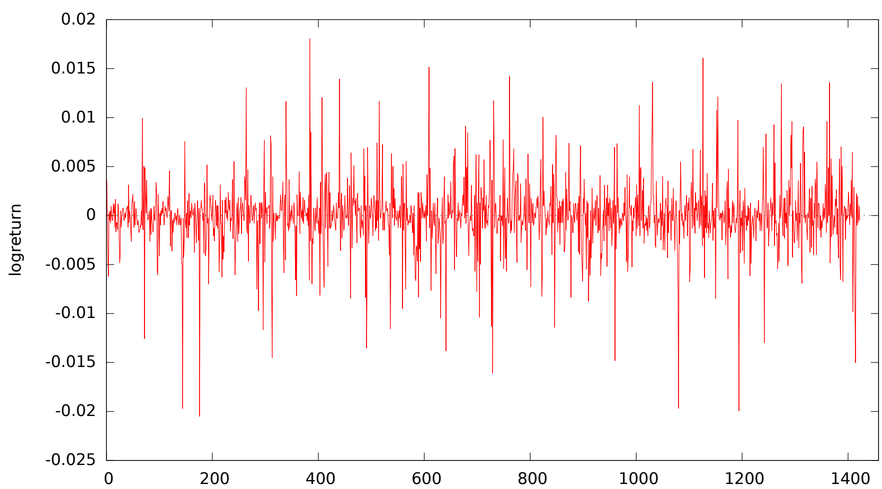
Mean	-8.3152e-005
Median	0.00000
Minimum	-0.012040
Maximum	0.011353
Standard deviation	0.0025362
C.V.	30.501
Skewness	-0.25569
Ex. kurtosis	3.0583
5% percentile	-0.0041567
95% percentile	0.0038373
Interquartile range	0.0025119

Source: author's computations.

4.8.3 NG

We also use Natural Gas NYMEX. Natural gas futures are the third-largest physical commodity futures contract in the world by volume. They are widely used as a national benchmark for the price of natural gas.

Figure 4.7: Log-returns of NG



Source: author's computations.

4.8.4 SV

Silver COMEX is one of the most important metal futures. Contract size is 5,000 troy ounces. Silver has a dual role, it is a precious metal for investment and also an industrial metal for commercial use. The price of this metal is

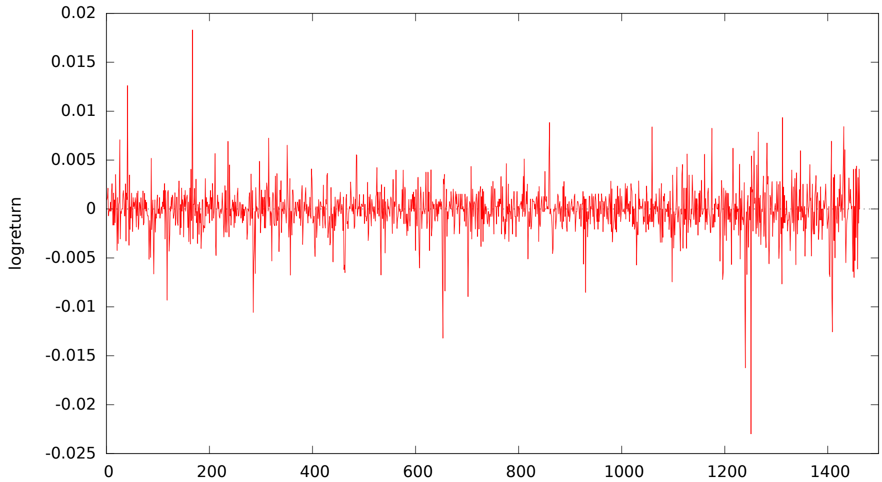
Table 4.6: Statistics of NG

Mean	-3.3771e-005
Median	0.00000
Minimum	-0.020508
Maximum	0.018080
Standard deviation	0.0035150
C.V.	104.08
Skewness	-0.31172
Ex. kurtosis	6.4224
5% percentile	-0.0055437
95% percentile	0.0054807
Interquartile range	0.0025195

Source: author's computations.

affected by mine production, industrial demand, and the general health of the world economy. The price of silver can be volatile beyond what many consider acceptable risk.

Figure 4.8: Log-returns of SV



Source: author's computations.

4.8.5 CL

Crude Oil Futures Contract are the world's most actively traded commodity based on crude oil. We have data from the New York Mercantile Exchange (NYMEX), which is the major trading exchange for crude oil futures contracts. The contract unit of Crude Oil Futures Contract is 1,000 barrels. Trading starts

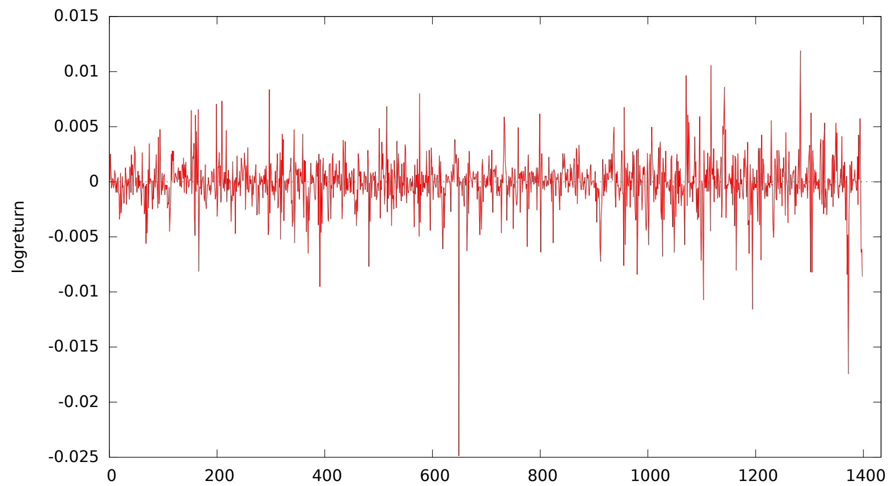
Table 4.7: Statistics of SV

Mean	-0.00014725
Median	0.00000
Minimum	-0.022996
Maximum	0.018314
Standard deviation	0.0024015
C.V.	16.309
Skewness	-0.72225
Ex. kurtosis	11.964
5% percentile	-0.0037173
95% percentile	0.0032493
Interquartile range	0.0022060

Source: author's computations.

every Sunday - Friday at 5:00 p.m. till 4:15 p.m. Chicago Time with a 45-minute break each day beginning at 4:15 p.m. Chicago Time.

Figure 4.9: Log-returns of CL



Source: author's computations.

4.8.6 CD

The contract size of CAD/USD futures is 100,000 Canadian dollars. CAD/USD futures are traded on Chicago Mercantile Exchange. The Canadian Dollar is the seventh most traded currency on the FOREX market. This currency pair has an advantage for us because of the English language in both countries as well as similar time zones, if we assume that it is traded mostly by domestic

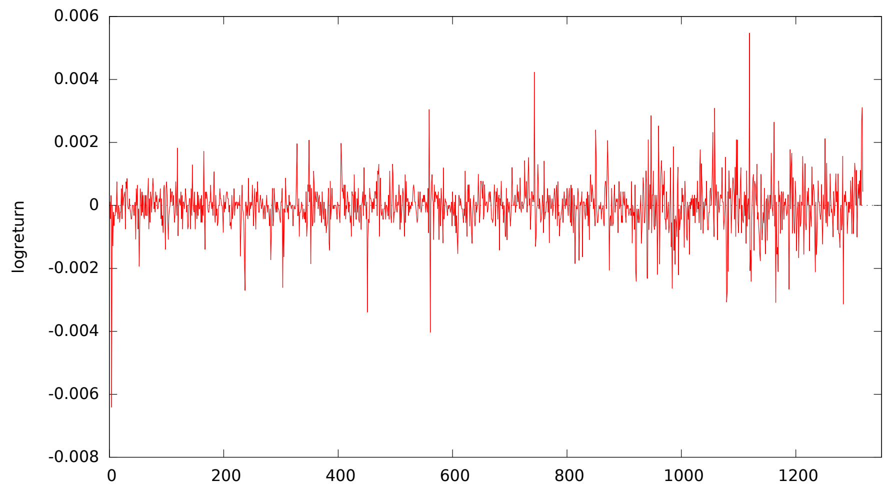
Table 4.8: Statistics of CL

Mean	-9.2847e-005
Median	0.00000
Minimum	-0.024885
Maximum	0.011892
Standard deviation	0.0023417
C.V.	25.221
Skewness	-1.2798
Ex. kurtosis	13.689
5% percentile	-0.0037343
95% percentile	0.0031035
Interquartile range	0.0019161

Source: author's computations.

investors. On the chart below, we see that log-returns of CD are more volatile in the second half of the period.

Figure 4.10: Log-returns of CD



Source: author's computations.

Table 4.9: Statistics of CD

Mean	-3.1383e-005
Median	0.00000
Minimum	-0.0064116
Maximum	0.0054783
Standard deviation	0.00076145
C.V.	24.263
Skewness	-0.27903
Ex. kurtosis	9.0934
5% percentile	-0.0011987
95% percentile	0.0010943
Interquartile range	0.00065850

Source: author's computations.

Chapter 5

Empirical results

In this chapter we present our empirical results. We comment on the results about the impact of Twitter sentiment on securities' returns and volatility. We begin with a Granger causality analysis, ARIMA models, Conditional Heteroscedastic models and realized volatility models. We discuss and analyze in-sample forecasts quality of the models. We also test three hypotheses exploring the impact of Twitter data on the financial markets.

Let us remind that we use only hourly frequency of the data in this work, so if we mention e.g. weekly log-differenced values of a particular Twitter sentiment variable, we mean the output from the equation 4.1.

5.1 Granger causality analysis

Granger causality test provides us information on whether Twitter-based variables can improve forecasts of securities' returns. GCA helps us to explore the significance and predictive power of Twitter-based lagged variables. It also enables us to test the hypotheses stated in chapter three. Based on the results of Granger causality tests shown in the tables 5.1 and 5.2, we can reject the null hypotheses at 1% level of significance that:

1. Twitter sentiment has no impact on security returns (i.e. changes in mood do not affect investors' decision making processes in financial markets),
2. Twitter keywords related to examined securities have no predictability power on log-returns.

From the results we see that examined securities' returns are influenced by very different emotions. We can illustrate it on example of two exchange rates on FOREX. The base currency in both currency pairs is USD, while quote currency is JY and CD, respectively. The most significant causal variable for JY is happiness, while for CD the variable happiness is insignificant in predicting the log-returns for all tested lags (i.e. from 1 to 7 hours). In case of JY, emotions confusion and caring have significant causal relations with the log-returns of JY. Concretely, confusion is significant at 10% level of significance for lags ranging from 3 to 7 hours, while caring for lags 1 and 2.

What interesting is that positive mood (its weekly log-differenced values) exhibit much bigger predictive power than mood negative (its weekly log-differenced values), which is at all lags 1-7 insignificant. When referring to the leverage effect, the impact of negative information (sentiment) should be greater than the impact of positive information (sentiment).

The keywords related to the symbols are Granger-causative of their log-returns. Past lag $l = 1$ (where the p-value equals 1.7%) the Granger causal relation between the keyword “CAD” and the log-returns becomes insignificant. The significant keyword for the causality with log-returns of JY is “USD” (its weekly log-differenced values), which exhibit, for lags ranging from 4 to 7 hours, a statistically significant correlation with the log-returns. For comparison, we add the performance of the daily log-differenced values of USD, which shows between lags 2 and 7 weaker predictive power.

Anger is, among psychologists, e.g. Goleman (1996), considered to be the strongest and most persistent emotion. Therefore we should be able to detect it in our data. For JY anger is only significant (at 10% level of significance) at lags 4, 5 and 7, but is insignificant at 5% level of significance. Anger shows some causality to the log-returns of NG, shown in table 5.2.

In the table 5.2, we show the most causal exogenous explanatory variables to the log-returns of several examined securities. For CL the most causal variable, according to the GCA, is the fear (its weekly differenced value). The keyword “silver” (its weekly differenced value) related to SV exhibits high predictive power. In particular, for lags ranging from 2 to 7 hours the keyword “silver” has highly significant causal relations with the log-returns (at e.g. 0.05% level of significance). The emotion afraid (its daily differenced values), from the lexicon based on Staiano & Guerini (2014), exhibits the Granger causal relation with CL’s log-returns. The causality decreases from lags 1 to 4 and is insignificant for the rest of the lags.

From Behavioral finance we know that anything affecting human emotions and mood can make systematic errors in investors’ decisions. Our results from

Table 5.1: GCA - statistical significance of Japanese Yen's log-returns at lags 1-7 hours (p-value < 0.01: ‡, p-value < 0.05: †, p-value < 0.1: *)

lag	mood positive(1D)	caring(1W)	USD(1D)	USD(1W)	happiness(1W)	anger(1W)	confusion(1W)
1	**0.0676***	**0.05***	0.19	0.9209	0.631	0.571	0.548
2	**0.0358†**	**0.0942***	**0.0857***	**0.0258†**	0.7967	0.3659	0.4392
3	0.1181	0.1707	**0.0256†**	**0.0053‡**	**0.0957***	0.1171	**0.0812***
4	0.2869	0.3004	**0.0538***	**0.0018‡**	**0.0055‡**	**0.0751***	**0.06***
5	0.5257	0.4337	**0.0928***	**0.0017‡**	**0.0056‡**	**0.0944***	**0.0657***
6	0.4195	0.4807	0.1079	**0.0033‡**	**0.0067‡**	0.1267	0**.0938***
7	0.7631	0.6147	0.2506	**0.0043‡**	**0.0031‡**	**0.0713***	**0.0603***

*** (1D) means daily log-differenced values of a particular Twitter sentiment, while (1W) means weekly log-differenced values of a particular Twitter sentiment*

Source: author's computations.

the Granger causality analysis based on Twitter sentiment data suggests that investors are not perfect homo economicus.

Table 5.2: GCA - statistical significance of securities' log-returns at lags 1-7 hours (p-value < 0.01: ‡, p-value < 0.05: †, p-value < 0.1: *)

lag	CL fear(1W)	CD afraid(1D)	S&P confusion(1W)	SV silver(1D)	NG anger(1D)
1	0.126	**0.002‡**	0.545	0.3488	0.645
2	**0.0755***	**0.0055‡**	**0.0135†**	**0.0003‡**	0.4969
3	**0.0222†**	**0.0166†**	**0.043†**	**0.0004‡**	**0.0579***
4	**0.0316†**	**0.0786***	**0.0987***	**8.75E-06‡**	0.1247
5	**0.0382†**	0.1682	0.1553	**1.96E-05‡**	0.195
6	**0.0496†**	0.2295	0.1186	**2.36E-05‡**	0.3017
7	**0.0155†**	0.304	**0.0906***	**0.0001‡**	0.2809

*** (1D) means daily log-differences, (1W) means weekly log-differences*

Source: author's computations.

5.2 Modelling returns

In this section we briefly examine the impact of the Twitter sentiment variables on log-returns, using ARIMA models with exogenous explanatory variables. We also explore the stability of the models.

JY

We use ARIMA models and the Box-Jenkins methodology to find the best model with an exogenous explanatory variable. As an exogenous explanatory variable, we use the daily logarithmic difference of the tag USD and the weekly logarithmic difference of negative mood. In table 5.3 we compare the models having an exogenous explanatory variable with the benchmark AR(1) model.

Based on the log-likelihood ratio test, it is obvious that the tag USD significantly improves the explanatory power of the model. The USD tag is significant; on the whole sample its coefficient equals to 0.0002907. The effect of the explanatory variable is quite small and is positive, so with increase in the tag's USD frequency we expect the JPY returns to rise. When running models on the smaller sample having 200 observations, the coefficient's value is almost always above zero and changes its value over time. It is seen in the figure 5.1, where we show the stability of the model.

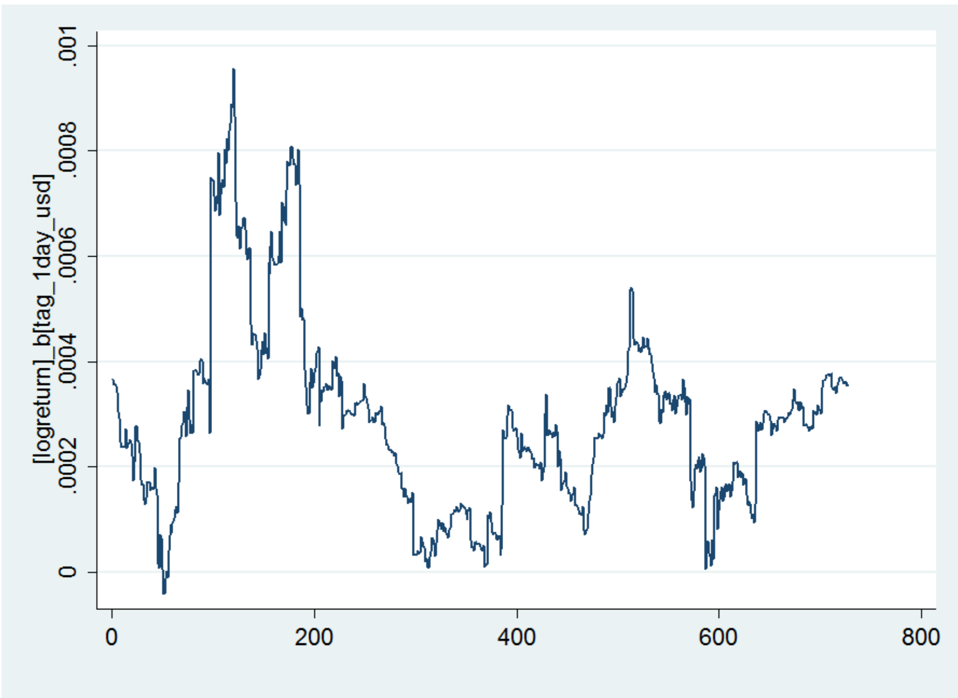


Figure 5.1: Stability of the AR(1) model for JY returns, where the explanatory variable is the USD tag

Table 5.3: Modelling JY returns using AR(1)

	plain AR(1) model			*model with tag USD*			*model with mood negative*		
Variable	Coefficient	Std. errors	P-value	Coefficient	Std. errors	P-value	Coefficient	Std. errors	P-value
AR	.0582837	0.0264727	0.028	0.0562522	0.0266388	0.035	0.0571679	0.0265559	0.031
constant	-.0000532	.000021	0.011	-0.0000509	0.0000209	0.015	-0.0000543	0.0000209	0.009
USD tag*	—	—	—	0.0002907	0.0001469	0.048	—	—	—
negative mood**	—	—	—	—	—	—	-0.0001741	0.0000772	0.024
sigma	0.0005994	7.38e-06	0.000	0.0005976	7.55e-06	0.000	0.0005974	7.59e-06	0.000
LL	5553.213			5559.437			5559.683		
BIC	-11079.11			-11091.55			-11092.04		
LLR test	—			0.0019813			0.00154923		

**The AR(1) model with the explanatory variable USD tag (daily logarithmic differences); **The AR(1) model with the explanatory variable negative mood (weekly logarithmic differences)*
LLR test stands for log-likelihood ratio test

Source: author's computations.

We conclude that the model is not very stable. This is not very surprising and we intuitively expected it to happen. Since we have measured just the total amount of the keyword USD, it does not reflect whether the U.S. dollar strengthened or weakened. But, based on the rolling analysis, we can conclude that most of the time were increases in frequency of USD tag associated with positive returns of JY.

The negative mood outperforms the plain AR(1) model and as the results of the log-likelihood ratio test suggests, adding negative mood into AR(1) model significantly improves explanatory power of the model.

In the figure 5.2 below, we show the stability of the model. Rolling uses again windows having 200 observations. The coefficient of the variable negative mood is quite unstable over time. Its value is negative in the whole period. When people (investors) feel bad, then they tend to be skeptical about the U.S. dollar's future and returns of JY go down and vice versa. We cannot forget that we track just English written tweets and we probably only accurately capture the mood of the USA and perhaps, English speaking investors, while Japanese mood is for us beyond the horizon.

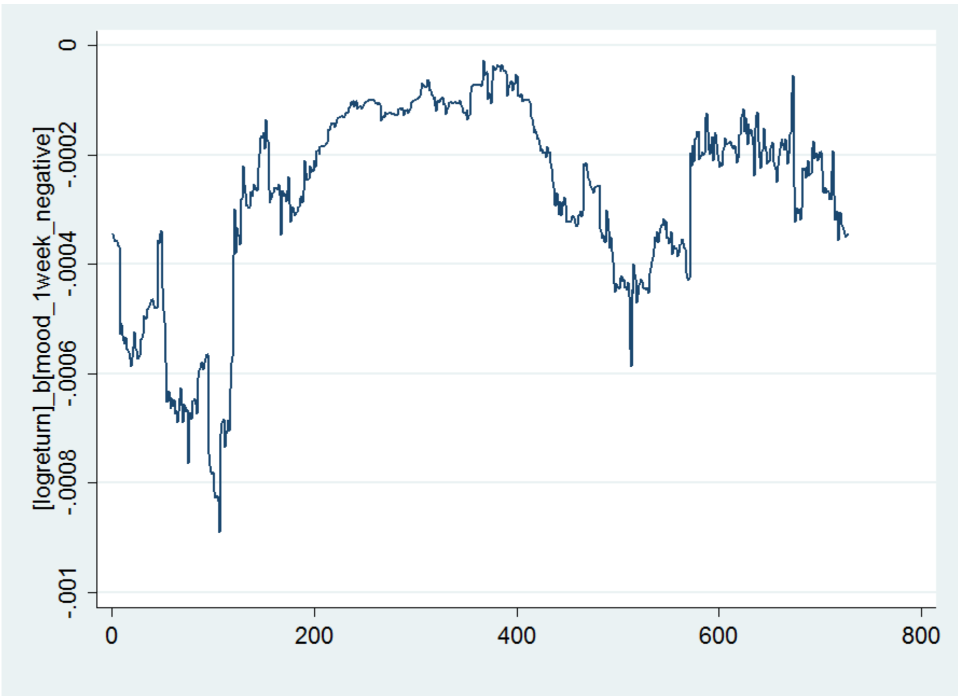


Figure 5.2: Stability of the AR(1) model for JY returns, where the exogenous explanatory variable is negative mood (weekly logarithmic differenced values).

5.3 Volatility estimation

In this section we introduce two types of models, the Conditional Volatility and Realized Volatility models.

5.3.1 Conditional Volatility models

We expect our data to better perform in volatility models rather than in models predicting returns. It is related to the nature of our Twitter variables, which tend to easily capture volatility, rather than the positive and negative changes in securities' returns. For our analysis, the GARCH(1,1) model is selected as the benchmark model. We add an explanatory variable into the variance equation of the GARCH model and compare the models' performance.

S&P

In the table 5.4 we present the performance of particular Twitter-based variables, which are separately put into the variance equation of the GARCH(1,1) model. In terms of explanatory power, the best performance among the emotions for S&P have happiness, fear, depression, afraid, amused, angry and annoyed. All these emotions are highly significant in the model, even at 0.1% level of significance. Moreover, the results of log-likelihood ratio test suggest that the models with these emotions are significantly better than the benchmark plain GARCH(1,1) model. Variables with the highest impact on the volatility, according to their coefficients' values, are afraid (-23.18) and angry (20.55). According to our model, when people (investors) are feeling afraid, the market is less volatile. The same holds for emotions caring and happiness, both emotions are related more to stability rather than dynamic changes. In cases where people (investors) are angry, the volatility of S&P index is higher. Similarly, when people (investors) are annoyed or amused, S&P index is more volatile. Both emotions amused and annoyed are related to strong affection, which can be perceived as a driver for changes, therefore there is a positive influence on volatility. The tag NASDAQ is also highly significant and has a positive effect on volatility of S&P index. The positive impact of NASDAQ tag is not surprising and goes along with intuition. S&P index comprises stocks listed on NASDAQ and when there is an increase in the tag NASDAQ in tweets, then there is probably something happening in either way to the market and S&P index. Positive tone is also highly significant and outperforms the benchmark model. The variable positive tone causes decrease in the volatility of S&P returns.

In conclusion of this subsection, we can reject the null hypothesis, stated in chapter three, that Twitter sentiment has no impact on volatility of security returns. We have found many variables strongly influencing volatility at 0.1% level of significance, e.g. emotions: happiness, fear, depression, afraid, amused, angry and annoyed.

Table 5.4: Explanatory power of the hourly logarithmic differenced variables for S&P volatility

*Variable**	Coefficient (Std. errors)	P-value	LL	BIC	LLR test
plain	—		2337.84	-4650.798	
afraid	-23.17733 (1.0684)	0.00001	2346.193	-4661.283	0.00023569
happiness	-16.03275 (2.169)	0.00001	2349.084	-4667.066	0.00001309
caring	-15.76263 (1.991345)	0.00001	2350.311	-4669.52	3.836e-06
positive tone	-13.71214 (1.539168)	0.00001	2348.13	-4665.156	0.00003397
tag NASDAQ	0.8739682 (0.1336886)	0.00001	2341.445	-4651.787	0.02718744
amused	18.6379 (1.05789)	0.00001	2348.107	-4665.11	0.00003476
annoyed	20.26948 (1.069633)	0.00001	2348.518	-4665.933	0.00002305
angry	20.55151 (1.068639)	0.00001	2348.468	-4665.832	0.00002423

**The explanatory variable in the variance equation of the GARCH(1,1) model*

Source: author's computations.

JPY

Based on the log-likelihood ratio test, it is clear that all variables in the table 5.5 significantly improve the explanatory power of the model. In terms of explanatory power, the best performance among the emotions has loneliness. All emotions in the table 5.5 are negative, meaning that the more the level of emotion increases, the less volatile JPY is. Negative mood is also highly significant and outperforms the benchmark model. In the figure 5.3, we plot the realized volatility and the predicted conditional volatility from our model with the lagged negative mood by one hour as an explanatory variable. Several large jumps in realized volatility are well captured by the predicted conditional volatility, indicating that the lagged independent variable has good explanatory power. In the FOREX market it is very difficult to predict jumps without tracking released news and information, because they are the key drivers there.

Table 5.5: Explanatory power of the hourly logarithmic differenced variables for JPY volatility

Variable*	Coefficient (Std. errors)	P-value	LL	BIC	LLR test
plain	—		8204.113	-16386.52	
mood negative	-4.0113 (0.2748199)	0.00001	8229.779	-16423.37	7.135e-12
loneliness	-4.202234 (0.2355922)	0.00001	8216.581	-16396.98	3.848e-06
remorse	-3.499793 (0.2150716)	0.00001	8215.948	-16395.71	7.246e-06
confusion	-3.996893 (0.2148865)	0.00001	8213.107	-16390.03	0.000124
anger	-3.969028 (0.2199979)	0.00001	8213.548	-16390.91	0.00008

**The explanatory variable in the variance equation of the GARCH(1,1) model*

Source: author's computations.

CD

From the GCA we got a very significant variable - afraid (its daily logarithmic differences). We use this - one hour lagged variable in our models. We try to put the variable into both mean and variance equations (not at the same time) and compare the performance with the benchmark model. As we see in the table 5.6, the independent variable improves the exlanatory power in both cases. The effect on returns is negative and small. The higher the one hour lagged variable afraid is, the lower the returns are and vice versa. For the explanatory variable in the variance equation, we observe a negative impact on volatility. So when one hour lagged afraid variable goes up, then the volatility of CD decreases. The tag CAD is also significant in the variance equation of the GARCH(1,1) model. Its coefficient's value is positive and equals 0.94, so its effect goes along with the intuition. An increase in tag's USD frequency causes

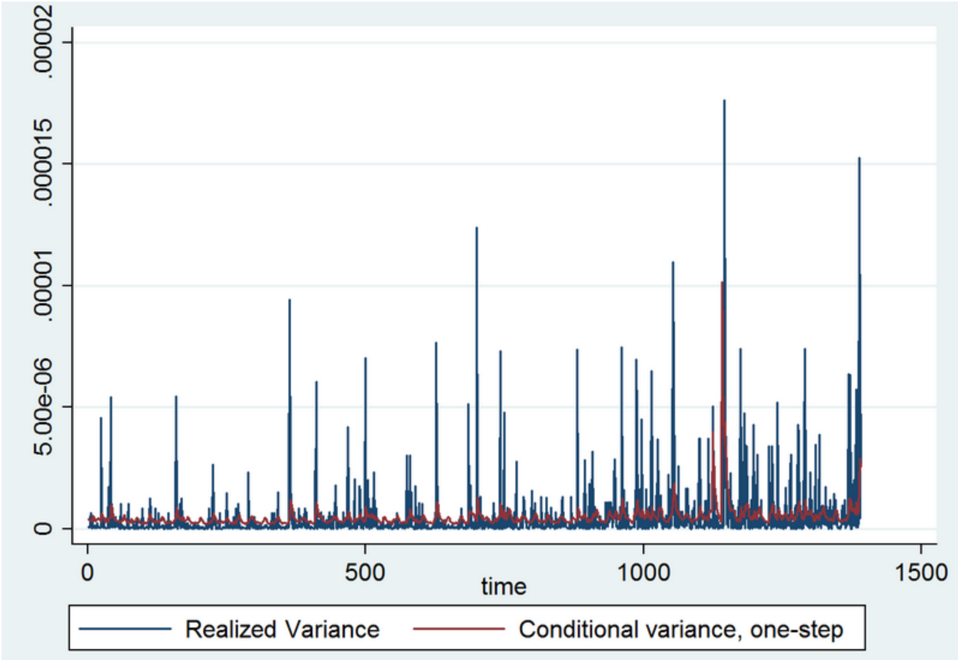


Figure 5.3: Plot of the predicted conditional volatility using one hour-lagged values of the negative mood and realized volatility of JPY.

a rise in volatility of CD returns and vice versa. Based on the log-likelihood ratio test we conclude that all variables in the table 5.6 significantly improve the explanatory power of the model.

SV

We again choose the variable for SV from the GCA, where we got a very significant variable - tag silver (hourly logarithmic differences). We use this - two hours lagged variable in our models. We try to put the variable into both mean and variance equation (not at the same time) and compare the performance with the benchmark model.

The explanatory variables are significant even at 0.1% level of significance. The variable tag silver in the mean equation underperforms the benchmark model, so in this case the explanatory variable did not improve the explanatory power of the model. On the other hand, the model with the explanatory variable tag silver in the variance equation outperforms the benchmark model considerably and, according to the result of log-likelihood test, the model with the explanatory variable significantly improves the explanatory power. The

Table 5.6: Explanatory power of the GARCH(1,1) model for CD

Variable	Coefficient (Std. errors)	P-value	LL	BIC	LLR test
plain	—	—	7630.959	-15233.19	
afraid*	-0.0003229 (0.0001376)	0.019	7636.893	-15237.87	0.00264787
afraid**	-0.8507895 (0.3340248)	0.011	7634.831	-15233.75	0.02081669
afraid***	-9.403759 (0.502727)	0.00001	7737.277	-15438.64	6.709e-47
tag CAD	0.9412258 (0.2031102)	0.000	7636.567	-15237.27	0.0036684

**The GARCH(1,1) model with the explanatory variable afraid (one hour lagged daily logarithmic difference) in the mean equation; **The GARCH(1,1) model with the explanatory variable afraid (one hour lagged daily logarithmic difference) in the variance equation; ***The GARCH(1,1) model with the explanatory variable afraid (one hour lagged hourly logarithmic difference) in the variance equation*

Source: author's computations.

Table 5.7: Explanatory power of the GARCH(1,1) model for SV

Variable	Coefficient (Std. errors)	P-value	LL	BIC	LLR test
plain	—		6804.725	-13580.31	
tag silver*	0.0007473 (0.0000628)	0.00001	6806.216	-13576	0.2251474
tag silver**	-.7303466 (0.0893997)	0.00001	6855.213	-13674	1.184e-22

**The GARCH(1,1) model with the explanatory variable tag silver (two hours lagged hourly logarithmic difference) in the mean equation; **The GARCH(1,1) model with the explanatory variable tag silver (two hours lagged hourly logarithmic difference) in the variance equation*

Source: author's computations.

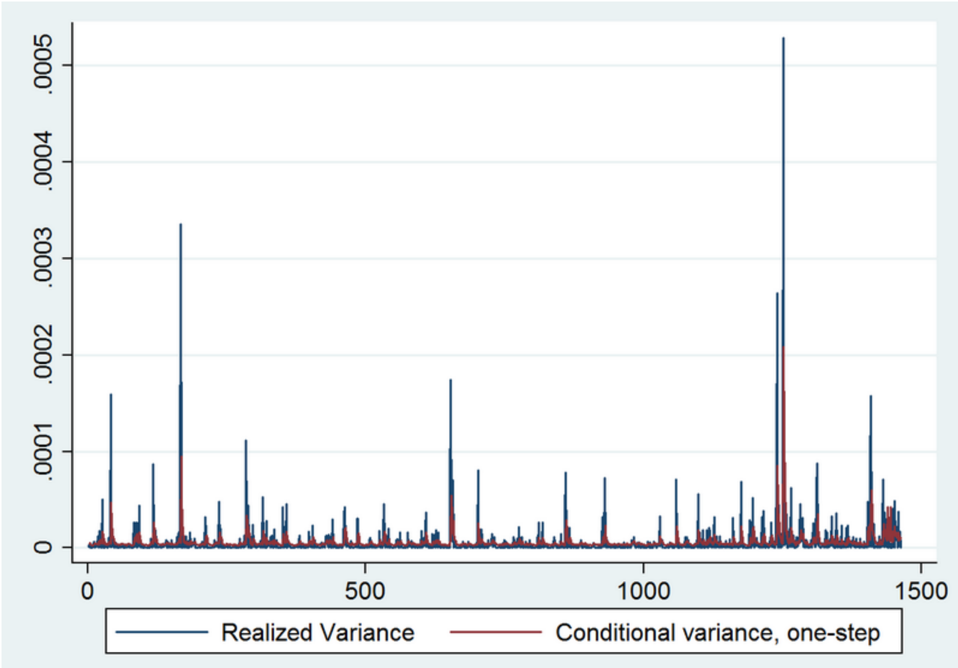


Figure 5.4: Volatility - realized and predicted conditional based on the model above for SV

volatility model performs really well, when comparing to realized volatility, we observe a good fit of the predicted conditional volatility. The effect of the two-hour lagged tag silver variable is negative. The more frequent the tag silver is in Tweets in past two hours(between two and three hours ago), the less volatile market is.

NG

Based on GCA, we choose the emotion anger for NG. We use the explanatory variable anger - three hours lagged hourly logarithmic difference. We try to put the variable into both mean and variance equation (not at the same time) and compare the performance with the benchmark model.

The explanatory variable in the mean equation is highly significant and by adding the variable the explanatory power of the model has significantly improved, as the log-likelihood ratio tests suggest. The variable in the variance equation outperforms the benchmark model and improves the explanatory power of the model. The volatility model performs really well, and when comparing to realized volatility we observe a good fit of the predicted conditional volatility. The effect of the three hours lagged anger variable is negative. The

Table 5.8: Explanatory power of the GARCH(1,1) model for NG

Variable	Coefficient (Std. errors)	P-value	LL	BIC	LLR test
plain GARCH(1,1)	—	—	6092.921	-12156.816	
anger*	-3.393485 (0.3727151)	0.00001	6103.384	-12170.48	0.000029
anger**	-4.287219 (0.2325751)	0.00001	6139.865	-12243.44	4.097e-21

**The GARCH(1,1) model with the explanatory variable anger (three hours lagged hourly logarithmic difference) in the mean equation; **The GARCH(1,1) model with the explanatory variable anger (three hours lagged hourly logarithmic difference) in the variance equation*

Source: author's computations.

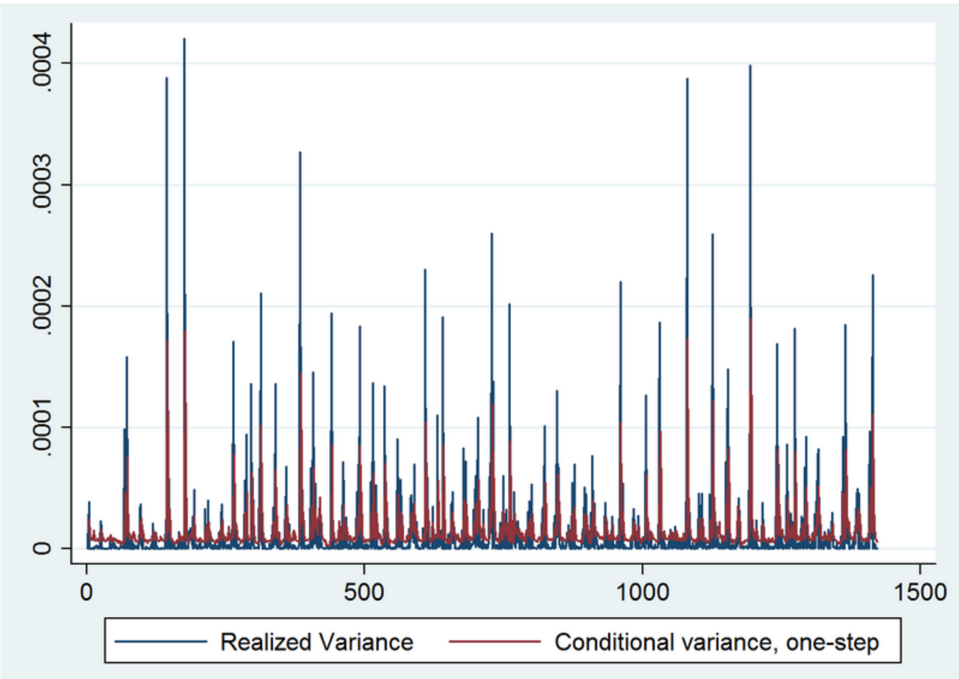


Figure 5.5: Volatility - realized and predicted conditional based on the model with explanatory variable anger (three hours lagged daily logarithmic difference) for NG

more angry people are in the three hours prior (between three and four hours ago), the less volatile market is. Using three-hour lagged variable may be interesting from the out-of-sample forecasting point of view.

5.3.2 Realized Volatility models

In this section we estimate Realized Volatility models. We introduce VAR model and HAR-RV model for S&P.

Firstly, we estimate a VAR model of order two. We use the model with an explanatory variable corresponding to monthly realized volatility of emotion caring (derived from hourly log-differenced values of the emotion caring). Daily and weekly variables are omitted due to lower significance. The VAR model's system of equations is defined:

$$\begin{aligned} & RV_t = \alpha_1 + \beta_{1,1} RV_{t-1} + \beta_{1,2} RV_{t-2} + \gamma_{1,1} EC_{t-1} + \\ & + \gamma_{1,2} EC_{t-2} + \delta_1 EC^{(22)}_{t-1} + \epsilon_{1,t} \\ & EC_t = \alpha_2 + \beta_{2,1} RV_{t-1} + \beta_{2,2} RV_{t-2} + \gamma_{2,1} EC_{t-1} + \\ & + \gamma_{2,2} EC_{t-2} + \delta_2 EC^{(22)}_{t-1} + \epsilon_{2,t} \end{aligned} \tag{5.1}$$

The output of the VAR model in shown in the figure 5.9. The results reveal significant autoregressive estimates for the realized volatility as well as for the emotion fear at lag one. The variable emotion fear is highly significant in the RV_t. The results of the Granger causality test suggests that the emotion fear provides significant information about future volatility.

HAR-RV model

In this section we use the HAR-RV model with three additional explanatory variables ERV, standing for realized variance of the hourly log-differenced values of a Twitter sentiment/tag variable (denoted ERV_{t-1}). We ran the model for all emotions based on Drummond (2004), tones and moods as well as for the tag NASDAQ. Let us remind the model's form:

$$\begin{aligned} & RV_t = \alpha_1 + \beta_1 RV_{t-1} + \beta_2 RV^{(5)}_{t-1} + \beta_3 RV^{(22)}_{t-1} + \gamma_1 ERV_{t-1} + \gamma_2 ERV^{(5)}_{t-1} \epsilon_{1,t} + \\ & + \gamma_3 ERV^{(22)}_{t-1} + \epsilon_t \end{aligned} \tag{5.2}$$

We provide the results in the tables 5.10 - 5.12, while test statistics on explanatory power and predictive accuracy are presented in tables A.6 - A.8.

Table 5.9: VAR Model - Estimation Results for S&P

	RV_t		EF_t	
Variable	Coef.	P-value	Coef.	P-value
RV_{t-1}	0.8436712	0.000	212.7181	0.209
	(0.0523341)		(169.3201)	
RV_{t-2}	0.0898342	0.086	-226.4183	0.181
	(0.0522864)		(169.166)	
EF_{t-1}	0.0000525	0.002	0.7256981	0.000
	(0.0000165)		(0.0535081)	
EF_{t-2}	-0.0000663	0.000	0.0761322	0.156
	(0.0000166)		(0.0536156)	
$EF_{t-1}^{(22)}$	0.0001885	0.015	0.1071671	0.668
	(0.0000773)		(0.2501932)	
$Constant$	-4.35e-06	0.037	0.0028499	0.672
	(2.08e-06)		(0.0067308)	
Granger causality test (p-value)	0.0000		0.5762	
R^2	0.8926		0.6245	
RMSE	2.6e-06		0.008563	

EF stands for realized variance of the emotion fear

Source: author's computations.

When checking the assumptions of the model, we observe heteroscedasticity in residuals, which is related to volatility of volatility. Therefore we need to apply heteroscedasticity-consistent standard errors; using Newey-West standard errors is not necessary, since we do not deal with autocorrelation.

Twitter based variables are mostly highly significant at longer horizon one month, while at daily horizon they are mostly insignificant. This holds for emotions loneliness, remorse, fear, depression, caring and happiness. Further, we point out that the models with Twitter based variables perform better than the plain model. When comparing the coefficient of determination of the plain model and the other models, we observe an improvement by values starting at 0.0019 to 0.0083.

The explanatory power of the models is evaluated using the log-likelihood ratio test. For all Twitter-based variables, apart from fear and variable afraid, presented in the tables 5.10 - 5.12 hold at 1% level of significance that adding them to the HAR-RV model significantly improves the models' explanatory power.

For analyzing the predictive accuracy, we use the Diebold-Mariano test, introduced by Diebold & Mariano (2002). The test compares the accuracy of two models based on the selected loss functions. The null hypothesis claims that two models have the same predictive accuracy. The best performance among the emotions have the emotions hurt and inadequateness. The test statictics of Diebold-Mariano test suggest that both emotions have significantly more accurate predictions than the HAR-RV model without any Twitter sentiment variable. Beside the emotions inadequateness and hurt, there is just one more variable having significantly more accurate predictions than the HAR-RV model, which is anger.

The coefficients' values of the variable standing for realized variance of the emotion are positive, so the bigger changes in an emotion the more volatile the returns are. Monthly variables are highly significant and have bigger impact on volatility than daily realized variances of the variables. These two emotions are also, beside the emotion confusion, the only emotions which are highly significant at daily horizons. The coefficients of daily emotions' variables have smaller impact of the volatility of S&P than monthly variables.

The emotion anger is insignificant in all horizons, therefore changes in anger (realized volatility of anger), however Diebold-Mariano test suggests that model's predictions are significantly more accurate than the plain HAR-RV model.

Emotions fear and afraid are outperformed by the other emotions. Also, positive mood has a bigger explanatory power than negative mood and both variables have positive influence on realized volatility. For the tag NASDAQ, there is a negative effect on the volatility in long-term (monthly data), while for weekly data there is a positive correlation with the volatility. This may suggest that the tag captures well the dynamic in price changes. The negative sign of the monthly coefficient refers to decrease in volatility when changes in the frequency of the tag NASDAQ goes up.

Based on our results and test statistics we conclude that using Twitter-based variables as an extension to classical HAR-RV model can significantly improve the explanatory power and predictive accuracy of the model and outperform the plain HAR-RV model.

Table 5.10: HAR-RV model

	Plain model		*happiness*		*Caring*		*Depression*		*Inadequateness*		*Fear*	
Variable	Coef.	P	Coef.	P	Coef.	P	Coef.	P	Coef.	P	Coef.	P
	Std. error		Std. error		Std. error		Std. error		Std. error		Std. error	
RV_{t-1}	0.7833998	0.000	0.781743	0.000	0.776998	0.000	0.773410	0.000	0.767843	0.000	0.760292	0.000
	0.0668368		0.065100		0.064839		0.065688		0.065541		0.069483	
$RV_{t-1}^{(5)}$	0.2067522	0.024	-0.310261	0.029	-0.291672	0.028	-0.267542	0.039	-0.241237	0.081	0.193993	0.027
	0.0908924		0.141613		0.131926		0.128900		0.137629		0.087246	
$RV_{t-1}^{(22)}$	0.1898075	0.143	1.308991	0.000	1.277736	0.000	1.230014	0.000	1.229798	0.000	0.242632	0.088
	0.1291961		0.331593		0.323319		0.30409		0.302572		0.141836	
ERV_{t-1}			0.000038	0.332	0.000041	0.295	0.000037	0.450	0.000089	0.010	0.000002	0.846
			0.000039		0.000039		0.000049		0.000034		0.000008	
$ERV_{t-1}^{(5)}$			0.000222	0.009	0.000244	0.021	0.000189	0.058	0.000117	0.079	0.000031	0.190
			0.000085		0.000105		0.000099		0.000067		0.000023	
$ERV_{t-1}^{(22)}$			0.000488	0.001	0.000410	0.005	0.000583	0.001	0.000208	0.014	0.000148	0.016
			0.000151		0.000144		0.000173		0.000085		0.000061	
Constant	-6.12E-07	0.219	-0.000014	0.000	-0.000013	0.000	-0.000014	0.000	-0.000010	0.000	-0.000006	0.004
	4.96E-07		0.000004		0.000003		0.000003		0.000003		0.000002	
R^2	0.8930		0.8986		0.8989		0.8989		0.9005		0.8949	
BIC	-7959.593		-7960.862		-7961.891		-7961.766		-7967.285		-7948.201	
LL	3991.507		4000.924		4001.428		4001.276		4004.185		3994.642	

EFRV stands for realized variance of the hourly log-differenced values of a variable

Source: author's computations.

Table 5.11: HAR-RV model

Variable	*Confusion* Coef. / Std. error	P	*Hurt* Coef. / Std. error	P	*Anger* Coef. / Std. error	P	*Loneliness* Coef. / Std. error	P	*Remorse* Coef. / Std. error	P	*Afraid* Coef. / Std. error	P
RV_{t-1}	0.771829	0.000	0.745843	0.000	0.789539	0.000	0.774915	0.000	0.787790	0.000	.7801755	0.000
	0.065765		0.067466		0.065440		0.067907		0.064815		.0375714	
$RV_{t-1}^{(5)}$	-0.172622	0.182	-0.145785	0.213	-0.054719	0.666	-0.173545	0.151	-0.156406	0.189	-.2918946	0.049
	0.129126		0.116967		0.126478		0.120500		0.118703		.1476176	
$RV_{t-1}^{(22)}$	1.087356	0.000	0.802148	0.000	0.696421	0.004	0.930203	0.001	0.943356	0.000	.9427529	0.000
	0.281729		0.215999		0.239661		0.266152		0.268232		.247128	
ERV_{t-1}	0.000093	0.010	0.000053	0.019	0.000034	0.250	-0.000019	0.638	0.000007	0.435	7.32e-08	0.998
	0.000036		0.000023		0.000029		0.000040		0.000009		.0000305	
$ERV_{t-1}^{(5)}$	0.000122	0.081	-0.000108	0.237	0.000075	0.322	0.000224	0.043	0.000037	0.086	-9.95e-06	0.870
	0.000070		0.000091		0.000075		0.000110		0.000021		.0000606	
$ERV_{t-1}^{(22)}$	0.000073	0.332	0.000652	0.000	0.000091	0.325	0.000321	0.044	0.000175	0.003	.0005772	0.000
	0.000075		0.000158		0.000092		0.000159		0.000058		.00016	
Constant	-0.000008	0.000	-0.000014	0.000	-0.000005	0.002	-0.000012	0.000	-0.000007	0.000	-9.95e-06	0.000
	0.000002		0.000003		0.000002		0.000003		0.000002		2.41e-06	
R^2	0.8997		0.9013		0.8959		0.8977		0.8969		0.8979	
BIC	-7964.668		-7970.276		-7951.638		-7957.801		-7955.11		-7935.104	
LL	4002.827		4005.631		3996.312		3999.393		3998.048		3988.035	

EFRV stands for realized variance of the hourly log-differenced values of a variable

Source: author's computations.

Table 5.12: HAR-RV model

Variable	*mood positive* Coef. / Std. error	P	*mood negative* Coef. / Std. error	P	*tone positive* Coef. / Std. error	P	*tone neutral* Coef. / Std. error	P	*tone negative* Coef. / Std. error	P	*tag Nasdaq* Coef. / Std. error	P
RV_{t-1}	0.76209	0.000	0.7629027	0.000	0.767502	0.000	0.758861	0.000	0.763308	0.000	0.759505	0.000
	0.064988		0.065905		0.066379		0.065918		6.63E-02		0.064992	
$RV_{t-1}^{(5)}$	-0.184243	0.118	-0.102370	0.341	-0.01240	0.906	-1.28E-01	0.225	-0.010503	0.919	-0.086637	0.607
	0.117492		0.1072594		0.1053751		0.10525		1.03E-01		0.1681787	
$RV_{t-1}^{(22)}$	1.28839	0.000	0.956826	0.000	0.707827	0.001	0.997849	0.000	0.673129	0.001	0.95715	0.005
	0.340393		0.2523406		0.213961		0.258595		0.1979925		0.3362078	
ERV_{t-1}	2.80E-06	0.283	7.43E-07	0.658	3.26E-06	0.124	-1.15E-06	0.732	2.97E-06	0.071	-1.66E-07	0.424
	2.61E-06		1.68E-06		2.12E-06		3.35E-06		1.64E-06		2.08E-07	
$ERV_{t-1}^{(5)}$	8.27E-06	0.337	4.27E-06	0.576	-3.82E-06	0.592	8.48E-06	0.459	-4.97E-06	0.493	3.06E-06	0.017
	8.60E-06		7.63E-06		7.12E-06		0.0000114		7.25E-06		1.27E-06	
$ERV_{t-1}^{(22)}$	0.000097	0.000	0.0000826	0.000	0.0000927	0.001	0.0001161	0.001	0.0000864	0.001	-0.000011	0.009
	0.000022		0.000022		0.000027		0.000035		0.000026		4.09E-06	
Constant	-7.02E-06	0.000	-4.98E-06	0.000	-4.26E-06	0.000	-0.000011	0.000	-3.66E-06	0.000	-1.04E-06	0.102
	1.77E-06		1.25E-06		1.14E-06		2.55E-06		9.65E-07		6.32E-07	
R^2	0.8996		0.8984		0.8971		0.8982		0.8979		0.8978	
BIC	-7964.26		-7959.987		-7955.648		-7959.493		-7957.021		-7958.094	
LL	4002.623		4000.486		3998.317		4000.239		3999.003		3999.54	

EFRV stands for realized variance of the hourly log-differenced values of a variable

Source: author's computations.

Chapter 6

Explanatory power of the variables using Wavelet Coherence

This chapter is dedicated to exploring the explanatory power of the variables using wavelet coherence. The wavelet coherence enables us to study the dependencies between two time series over time across different frequencies. The spearhead of using wavelets in economics and finance James B. Ramsey emphasizes their importance as a mighty tool for the explanatory power analysis, e.g. in Ramsey (2002).

6.1 JY

In the figure 6.1 we can see the estimated wavelet coherence and the phase difference for selected Twitter sentiment variables (listed as the first variable in the wavelet coherence) and log-return of JY from scale 1 hour to 512 hours, which is approximately 1 trading month. The scale is measured on the vertical axis, while time on the horizontal axis. The color shows the measure of dependency between the time series. The areas bordered by black lines depict significant dependence, at the 5% significance level, which was assessed using Monte Carlo simulations.

The first picture in Figure 6.1 examines both the frequency bands and time intervals where the mood positive (measured as daily logarithmic difference) and log-return of JY move together. There are small regions of significant correlation at lower scales (i.e. higher frequencies), for instance between 1 and 12 hours, throughout almost the whole period. Between September 8 and September 22 we observe a large area of highly significant local correlations at scales from 1 day up to 2 days. In this region, twitter sentiment (positive mood) seems to lead market data and the correlation is negative. In the period between September 4 and September 22 there is a negative correlation at scales from 2 days up to 54 hours and the leading time series are JY log-returns.

On the lower frequencies around 1-2 market weeks we can see that JY log-returns negatively influence mood positive. The significant correlation persists over all examined periods. The large time-scale area on the lowest frequencies between 2 and 4 market weeks reveals a dependency as well, while the leading series are JY log-returns. Emotion caring leads the log-return of JY in the period between August 27 and October 2 at scales around 1 trading week. The USD keyword variables (daily and weekly log-differenced values of number of tweets containing the keyword USD per hour) has significant local correlations with JY log-returns on small areas at scales from 1 hour up to 1 day. USD keyword variables influence the JY log-returns between August 27 and September 9 at scales 40-60 hours. The dependency between emotions and JY log-returns appears on small areas at scales 1-6 hours and 12-48 hours. Between July 15 and July 21, the JY log-returns negatively influence the emotions confusion and anger (weekly log-differenced values).

6.2 S&P

In Figures 6.2 - 6.4 we show the wavelet coherence of S&P for all examined emotions based on both Staiano & Guerini (2014) and Drummond (2004). The first variables in the figures are emotions, while the second are log-returns. The results for emotions, based on Drummond (2004), are in the long-term positively correlated with S&P log-returns; emotions are the leading variable. There are some notable local correlations at lower scales. Between August 8-25, 2014 we observe a negative correlation between caring and log-returns as well as depression and log-returns, where the S&P log-returns are the leader in the relationship. At scales around 1 trading week between August 25 and September 9 the emotion remorse is positively influenced by the S&P log-returns. The S&P log-returns lead emotions caring, depression, fear and remorse between July 30 and August 11, and there is a negative correlation.

Emotions based on Staiano & Guerini (2014) behaves similarly. At the lowest frequencies there is a positive correlation with the S&P close price, which is led by emotions. At lower scales, up to 1 trading day, we mostly observe small local correlations throughout the whole analyzed period. The

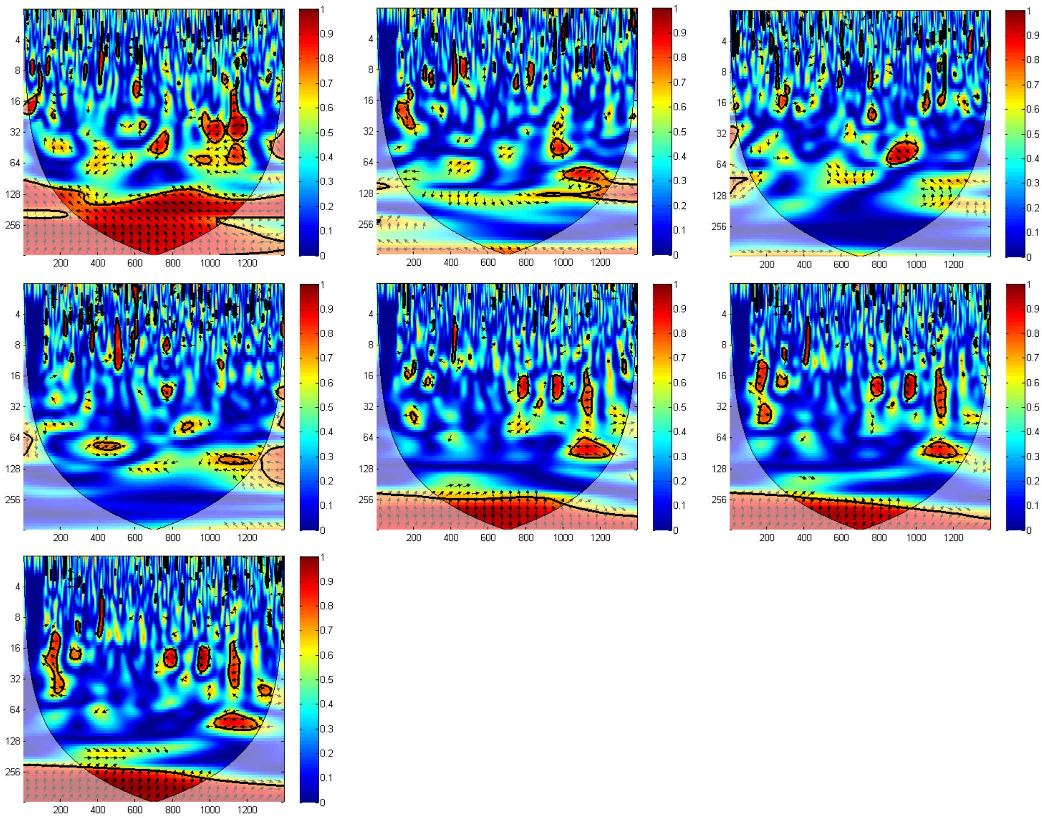


Figure 6.1: Wavelet coherence of JY in pairs with the same Twitter sentiment as used in the table 5.1

**Twitter sentiment variables used: mood positive (1D), caring (1W), USD (1D), USD (1W), happiness (1W), anger (1W) and confusion (1W). The scale is measured on the vertical axis, while time on the horizontal axis. The color shows the measure of dependency between the time series. The black lined regions depict significant dependence, at the 5% significance level, which was assessed using Monte Carlo simulations.*

*** (1D) means daily log-differences, (1W) means weekly log-differences* notable period of negative correlation is between July 31 and August 19 at scales around 2 trading days, where price leads the emotions. This area of dependency is the largest by the emotion afraid. The period of August 8-25 reveals a negative correlation between scales of 2 and 5 trading days, where at higher scales emotions lead the price, while at lower scales there is no leader in the relationship. All of these plots of emotions based on Staiano & Guerini (2014) are quite similar, they differ in size of particular areas on the same time-scale position. The emotion with the largest impact on emotions seems to be afraid and anger.

We can summarize the results from the wavelet coherence between emotions and S&P market data that at scales around one trading month there is a statistically significant positive correlation, in which emotions lead market data. This suggests that emotions influence investors' decision making processes at American stock markets (NYSE and NASDAQ) in the long run.

In Figure 6.4 we show the wavelet coherence of moods (positive and negative) and S&P log-returns. In the first plot at scales between 2 and 5 trading days we observe a large area of local negative correlation between positive mood and S&P log-returns. The correlation is present between July 17 and August 12 and the leader in the relationship are the S&P log-returns.

Both moods positively influence S&P log-returns at higher scales around 1 trading month throughout the whole analyzed period.

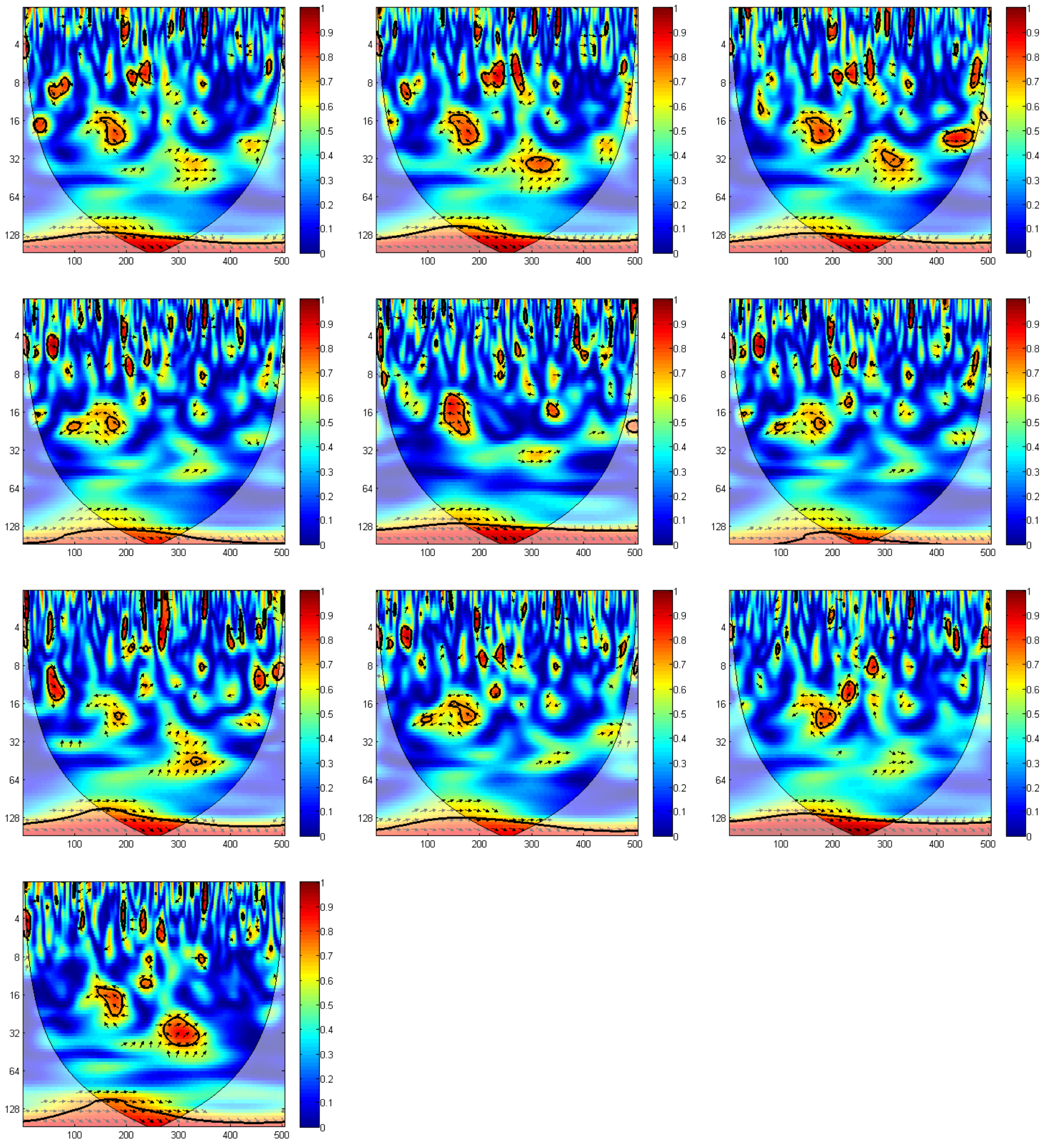


Figure 6.2: Wavelet coherence of S&P log-returns (listed as the second variable) in pairs with emotions based on Drummond (2004)

**Twitter-based variables of emotions used are the daily values of emotions: happiness, caring, depression, inadequateness, fear, confusion, hurt, anger, loneliness and remorse. The scale is measured on the vertical axis, while time on the horizontal axis. The color shows the measure of dependency between the time series. The black lined regions depict significant dependence, at the 5% significance level, which was assessed using Monte Carlo simulations.*

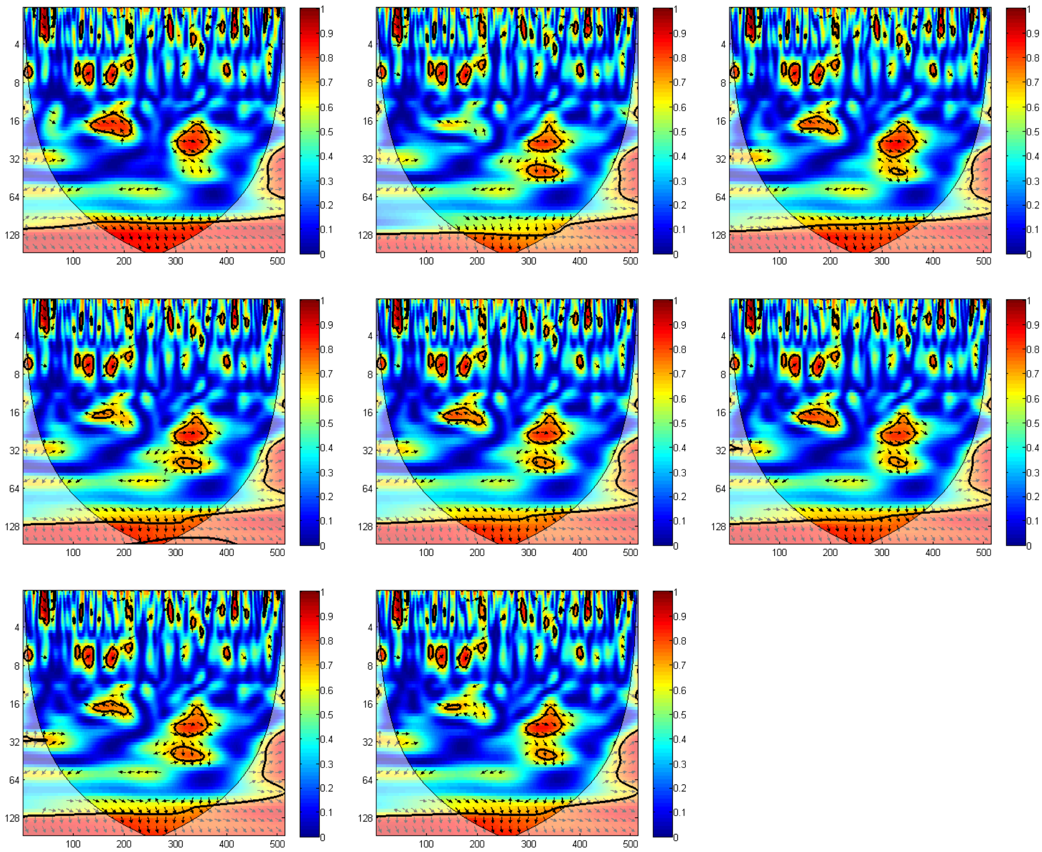


Figure 6.3: Wavelet coherence of S&P close price (listed as the second variable) in pairs with emotions based Staiano & Guerini (2014)

**Twitter-based variables of emotions used are the hourly values of emotions: afraid, amused, angry, annoyed, don't care, happy, inspired and sad. The scale is measured on the vertical axis, while time on the horizontal axis. The color shows the measure of dependency between the time series. The black lined regions depict significant dependence, at the 5% significance level, which was assessed using Monte Carlo simulations.*

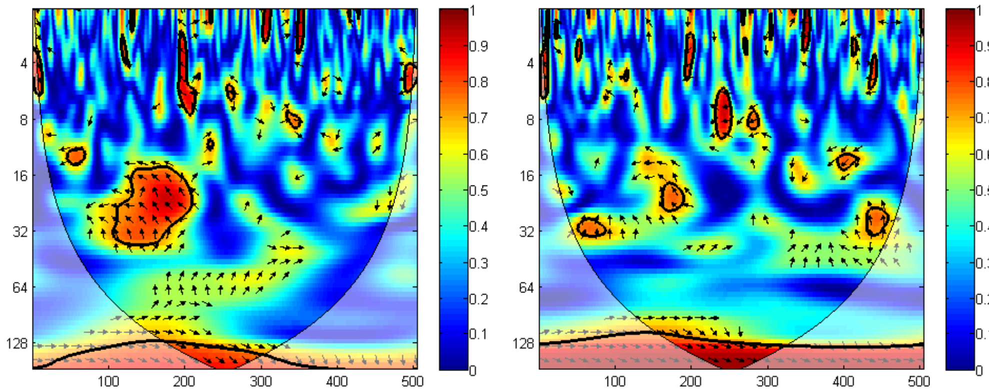


Figure 6.4: Wavelet coherence of S&P log-returns (listed as the second variable) in pairs with moods based Drummond (2004),

**Twitter-based variables used in this figure are logarithmic daily differences of both positive and negative moods. The scale is measured on the vertical axis, while time on the horizontal axis. The color shows the measure of dependency between the time series. The black lined regions depict significant dependence, at the 5% significance level, which was assessed using Monte Carlo simulations.*

Chapter 7

Conclusion

This work concentrates on the influence of social networks to financial markets. We introduced a novel approach to Twitter sentiment analysis, in which we collect continuous stream of data and analyze it. Our original data set contains over 200 million English written Tweets from the period between July 1, 2014 and October 9, 2014. On hourly data we investigate how are investors influenced by basic emotions, moods and sentiment in their decision making processes as well as the influence of keywords related to specific securities and FOREX symbols. We show that Twitter based variables may be an efficient tool when predicting securities' returns and volatility. Moreover, we reveal the influence of basic emotions on investors' decision making processes. The influence of basic emotions on investors differs for each security. In general, emotions fear and anger are very influential. Caring, happiness, inadequateness, hurt and annoyance have the biggest impact on S&P volatility. Emotions causing the biggest changes in volatility of JY returns are loneliness, confusion and remorse. Besides these emotions tag USD (for JY) and silver (for SV) can significantly improve the prediction accuracy as well as Twitter variable negative mood.

Particularly, we examine the relationships between Twitter-based variables and returns and volatility of several financial instruments. We explore the in-sample predictability of JY returns and discover that the tag USD and negative mood significantly improve the explanatory power of the models. Two AR(1) models with these exogenous explanatory variables outperform plain AR(1) model. The stability of the models indicates that coefficients of Twitter-based variables changes a lot over time, however, the coefficients' sign remain the same. The coefficient of the tag USD oscillates above zero, while the coefficient of the variable negative mood fluctuates below zero.

The influence of Twitter sentiment on volatility of securities' returns is tested and shown on both conditional and realized volatility models. In our analysis we provide results for JY, CD, NG, SV, CL and S&P. We discover that Twitter sentiment variables can significantly improve both explanatory power and predictive accuracy, which we have tested using the Diebold-Mariano test.

In order to investigate the explanatory power of Twitter-based variables, we apply wavelet coherence. Our results suggest that investors are influence by emotions and moods, especially at longer investment horizons. The impact of emotions at shorter investment horizons is limited and differs for particular securities as well as emotions.

Our recommendations on future follow-ups is, when using high frequency data up to hourly data, to focus on volatility prediction, for which the Twitter sentiment data are much more suitable than for predicting returns. The employment of realized volatility models and realized volatilities of Twitter sentiment variables seems to bring promising results. In case of larger data set, wavelet coherence analysis may be also a very efficient tool for exploring the explanatory power of Twitter-based variables.

In conclusion, our results can be used especially for prediction of both conditional and realized volatility. Twitter sentiment variables contain additional information about volatility of a particular financial instrument and have the ability to improve the model's performance. Predicting volatility is very important in financial markets, since risk is usually expressed by volatility. The ability to predict volatility is useful, for example, in risk management, asset pricing, hedging etc.

Bibliography

ADDISON, P. (2002): *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*. Taylor & Francis.

ANTWEILER, W. & M. Z. FRANK (2004): "Is all that talk just noise? the information content of internet stock message boards." *The Journal of Finance* **59(3)**: pp. 1259–1294.

BECHARA, A. & A. R. DAMASIO (2005): "The somatic marker hypothesis: A neural theory of economic decision." *Games and economic behavior* **52(2)**: pp. 336–372.

BOLLEN, J., H. MAO, & X. ZENG (2011): "Twitter mood predicts the stock market." *Journal of Computational Science* **2(1)**: pp. 1–8.

BOLLERSLEV, T. (1986): "Generalized autoregressive conditional heteroskedasticity." *Journal of econometrics* **31(3)**: pp. 307–327.

CHEN, H., P. DE, Y. J. HU, & B.-H. HWANG (2014): "Wisdom of crowds: The value of stock opinions transmitted through social media." *Review of Financial Studies* **27(5)**: pp. 1367–1403.

CHOWDHURY, S. G., S. ROUTH, & S. CHAKRABARTI (2014): "News analytics and sentiment analysis to predict stock price trends." *International Journal of Computer Science and Information Technologies* **5(3)**: pp. 3595–3604.

CORSI, F. (2004): "A simple long memory model of realized volatility." *Available at SSRN 626064* .

CORSI, F., S. MITTNIK, C. PIGORSCH, & U. PIGORSCH (2008): "The volatility of realized volatility." *Econometric Reviews* **27(1-3)**: pp. 46–78.

Diebold, F. X. & R. S. Mariano (2002): "Comparing predictive accuracy." *Journal of Business & economic statistics* **20(1)**.

Dimpfl, T. & S. Jank (2012): "Can internet search queries help to predict stock market volatility?" In "Paris December 2012 Finance Meeting EUROFIDAI-AFFI Paper," .

Drummond, T. (2004): "Vocabulary of Emotions." Available online `http://www.sba.pdx.edu/faculty/mblake/448/FeelingsList.pdf` accessed 15.6.2014.

Engle, R. F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation." *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Fama, E. F. (1970): "Efficient capital markets: A review of theory and empirical work*." *The journal of Finance* **25(2)**: pp. 383–417.

Ferguson, N., J. Guo, H. Lam, & D. Philip (2011): "Media sentiment and uk stock returns." *Working Paper* .

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, & N. A. Smith (2011): "Part-of-speech tagging for twitter: Annotation, features, and experiments." In "Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2," pp. 42–47. Association for Computational Linguistics.

Goleman, D. (1996): *Emotional Intelligence: Why It Can Matter More Than IQ*. Bantam Books.

Granger, C. W. (1969): "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica: Journal of the Econometric Society* pp. 424–438.

Grinsted, A., J. C. Moore, & S. Jevrejeva (2004): "Application of the cross wavelet transform and wavelet coherence to geophysical time series." *Nonlinear processes in geophysics* **11(5/6)**: pp. 561–566.

Hemalatha, G., G. S. Varma, & A. Govardhan (2013): "Sentiment analysis tool using machine learning algorithms." *International Journal of Emerg-*

ing Trends & Technology in Computer Science (IJETTCS) **2(2)**: pp. 105–109.

Karabulut, Y. (2013): "Can facebook predict stock market activity?" In "AFA 2013 San Diego Meetings Paper," .

Kramer, A. D., J. E. Guillory, & J. T. Hancock (2014): "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences* **111(24)**: pp. 8788–8790.

Medhat, W., A. Hassan, & H. Korashy (2014): "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* **5(4)**: pp. 1093–1113.

Müller, U. A., M. M. Dacorogna, R. Davé, O. V. Pictet, R. B. Olsen, & J. R. Ward (1993): "Fractals and intrinsic time: A challenge to econometricians." *Unpublished manuscript, Olsen & Associates, Zürich* .

Müller, U. A., M. M. Dacorogna, R. D. Davé, R. B. Olsen, O. V. Pictet, & J. E. von Weizsäcker (1997): "Volatilities of different time resolutionsâ€"analyzing the dynamics of market components." *Journal of Empirical Finance* **4(2)**: pp. 213–239.

Ramsey, J. B. (2002): "Wavelets in economics and finance: past and future." *Studies in Nonlinear Dynamics & Econometrics* **6(3)**.

Rao, T. & S. Srivastava (2012): "Analyzing stock market movements using twitter sentiment analysis." In "Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)," pp. 119–123. IEEE Computer Society.

Sakaki, T., M. Okazaki, & Y. Matsuo (2010): "Earthquake shakes twitter users: real-time event detection by social sensors." In "Proceedings of the 19th international conference on World wide web," pp. 851–860. ACM.

Scott, G. (2011): "Tone and Mood." Available online `http://www.sba.pdx.edu/faculty/mblake/448/FeelingsList.pdf` accessed 8.6.2014.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, & X. Deng (2013): "Exploiting topic based twitter sentiment for stock prediction." In "ACL (2)," pp. 24–29.

Soon, C. S., M. Brass, H.-J. Heinze, & J.-D. Haynes (2008): "Unconscious determinants of free decisions in the human brain." *Nature neuroscience* **11(5)**: pp. 543–545.

Sprenger, T. O., A. Tumasjan, P. G. Sandner, & I. M. Welpe (2014): "Tweets and trades: The information content of stock microblogs." *European Financial Management* **20(5)**: pp. 926–957.

Staiano, J. & M. Guerini (2014): "DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News." *Proceedings of ACL-2014* .

Tayal, D. & S. Komaragiri (2009): "Comparative analysis of the impact of blogging and micro-blogging on market performance." *International Journal* **1(3)**: pp. 176–182.

Torrence, C. & P. J. Webster (1999): "Interdecadal changes in the enso-monsoon system." *Journal of Climate* **12(8)**: pp. 2679–2690.

Tumarkin, R. & R. F. Whitelaw (2001): "News or noise? internet postings and stock prices." *Financial Analysts Journal* **57(3)**: pp. 41–51.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, & I. M. Welpe (2010): "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM* **10**: pp. 178–185.

Zusne, L. & W. H. Jones (2014): *Anomalistic psychology: A study of magical thinking.* Psychology Press.

Appendix A

Tables

Table A.1: Statistics of the emotion variables based on Drummond (2004)

Variable	Mean	Median	Min	Max	Std. Dev.	Skewness	Kurtosis
happiness	11389	12145	747.50	28126	2958.3	-0.15544	0.070386
caring	12158	12920	787.20	30642	3298.0	-0.063141	0.33877
depression	8341.7	8855.8	544.00	22565	2106.6	0.0011575	1.2301
inadequateness	2334.0	2411.6	154.60	8639.3	635.72	0.54978	5.5442
fear	226.83	231.40	14.100	833.20	71.613	1.4327	8.5101
confusion	3607.5	3722.3	244.60	14286	1001.0	0.76550	7.4867
hurt	765.60	801.70	52.600	2416.3	230.30	0.13057	1.3786
anger	2673.6	2770.4	174.00	9168.3	747.59	0.37655	3.4383
loneliness	2604.4	2800.3	158.00	6984.7	774.63	-0.0033827	0.68087
remorse	474.75	490.30	33.100	1451.5	144.70	0.24353	1.2328
happiness (H)	-0.00066058	0.009873	-1.0411	1.1863	0.12812	-0.06591	10.127
caring (H)	-0.00098356	0.012798	-1.0512	1.1892	0.13075	-0.11618	10.441
depression (H)	-0.00068118	0.009205	-1.0592	1.1902	0.12745	0.0051805	11.729
inadequateness (H)	-0.0012755	0.009888	-1.1465	1.2466	0.15417	0.16428	9.8669
fear (H)	-0.00082399	0.006215	-1.0537	1.2635	0.19274	0.2065	5.0236
confusion (H)	-0.0012607	0.011801	-1.1448	1.23	0.15736	0.24029	9.5511
hurt (H)	-0.00088149	0.016696	-1.1743	1.3003	0.15496	-0.18573	7.7989
anger (H)	-0.0015531	0.012781	-1.1199	1.2393	0.15471	0.053893	8.3724
loneliness (H)	-0.0017418	0.013047	-1.096	1.2071	0.1464	-0.29035	7.6158
remorse (H)	-0.0026862	0.006777	-1.3834	1.2456	0.1722	-0.30355	7.1276
happiness (D)	0.014277	0.009588	-1.0752	2.5735	0.17104	3.3125	55.677
caring (D)	0.01271	0.004869	-1.0852	2.5988	0.17489	3.1499	53.527
depression (D)	0.012624	0.005373	-1.0644	2.6163	0.17025	3.5237	58.921
inadequateness (D)	0.0098331	0.005261	-1.1449	2.569	0.19668	2.2893	35.137
fear (D)	0.01046	0.006629	-1.1921	2.6689	0.24503	1.6989	18.471
confusion (D)	0.010401	0.004667	-1.1509	2.5484	0.20155	2.0749	31.454
hurt (D)	0.0067477	0.002135	-1.1941	2.5343	0.19438	2.0733	36.313
anger (D)	0.01051	0.005559	-1.168	2.5554	0.19746	2.2044	33.657
loneliness (D)	0.012035	0.002095	-1.1748	2.6618	0.18918	2.6854	46.054
remorse (D)	0.0021159	-0.001315	-1.4075	2.5009	0.21121	1.6952	26.682
happiness (W)	-0.010355	-0.001665	-1.0489	1.8691	0.2337	0.061812	6.7592
caring (W)	-0.011717	-0.003783	-1.0958	1.9002	0.2417	0.026899	6.9282
depression (W)	-0.010095	-0.00058	-1.0359	1.8197	0.22693	0.13041	6.8526
inadequateness (W)	-0.016815	-0.002798	-1.3069	1.8163	0.25529	-0.24137	6.9545
fear (W)	-0.012319	0	-1.3262	1.8262	0.29288	0.12686	4.4473
confusion (W)	-0.016529	-0.003536	-1.3418	1.7766	0.26012	-0.21801	6.5927
hurt (W)	-0.016745	-0.006211	-1.3664	2.0813	0.26942	-0.18313	7.2618
anger (W)	-0.014821	-0.000443	-1.3496	1.8081	0.2596	-0.19965	6.6181
loneliness (W)	-0.010556	-0.001828	-1.2248	1.9671	0.26785	0.056764	6.5513
remorse (W)	-0.016559	-0.001673	-1.3959	1.9091	0.28104	-0.13153	5.7232

(1H) means hourly log-differences, (1D) means daily log-differences, (1W) means weekly log-differences

Source: author's computations.

Table A.2: Statistics of tag variables

Variable	Mean	Median	Min	Max	Std. Dev.	Skewness	Kurtosis
tag oil	438.84	474	0	631	114.91	-1.7132	2.4228
tag silver	462.15	494	0	606	115.86	-1.8975	3.2342
tag nasdaq	222.08	209	3	557	121.68	0.49703	-0.64117
tag cad	361.55	380	0	556	100.06	-0.88786	0.60042
tag usd	401.04	430	0	566	102.72	-1.69	2.5825
tag oil (H)	-0.0029803	0	-6.1696	4.5951	0.35105	-3.7168	148.83
tag silver (H)	-0.0031904	0	-6.2146	4.6151	0.35581	-3.7695	147.7
tag nasdaq (H)	0.0061418	-0.020479	-2.7726	2.7726	0.44462	0.21041	7.9425
tag cad (H)	-0.00036897	0	-6.0039	4.8752	0.30795	-4.6492	181.44
tag usd (H)	-0.004962	-0.00221	-6.129	4.5644	0.35195	-3.6392	145.66
tag oil (D)	-0.0173	-0.004082	-5.6455	5.6664	0.37568	-2.3764	104.78
tag gas (D)	-0.012739	0	-5.989	6.2146	0.45223	-1.1928	103.15
tag nasdaq (D)	0.14753	0.06337	-4.0073	4.2305	0.69458	0.33373	4.9851
tag cad (D)	0.014211	0.010152	-5.7071	5.3613	0.35635	-2.0371	103.86
tag usd (D)	-0.024289	-0.011173	-5.6131	5.425	0.36932	-2.834	106.21
tag oil (W)	-0.099712	-0.008457	-6.089	5.2781	0.44957	-3.6472	66.807
tag silver (W)	-0.098719	-0.002002	-6.1506	5.2883	0.43306	-4.4157	81.015
tag nasdaq (W)	-0.0020056	0	-3.1676	4.287	0.72094	0.51826	4.6402
tag cad (W)	-0.058651	0	-6.157	5.2933	0.43544	-2.739	62.61
tag usd (W)	-0.094761	-0.013015	-6.142	5.1874	0.42698	-4.4934	84.021

(1H) means hourly log-differences, (1D) means daily log-differences, (1W) means weekly log-differences

Source: author's computations.

Table A.3: Stationarity tests

Log-returns	ADF	P-value	KPSS	P-value
happiness	-0.765226	0.385	1.33753	< 0.01
caring	-0.786901	0.3755	1.36686	< 0.01
depression	0.82114	0.3605	1.19043	< 0.01
inadequateness	-0.747894	0.3927	2.9165	< 0.01
fear	-0.667283	0.4284	1.17281	< 0.01
confusion	-0.74072	0.3959	3.02572	< 0.01
hurt	-0.895022	0.3286	2.17486	< 0.01
anger	-0.749193	0.3921	2.54944	< 0.01
loneliness	-0.714025	0.4077	0.874459	< 0.01
remorse	-0.978533	0.2939	1.37775	< 0.01
positive mood	-0.599271	0.4581	0.528753	0.040
negative mood	-0.671752	0.4264	0.0449135	< 0.01
positive tone	-0.739686	0.3963	1.98929	< 0.01
neutral tone	-0.726034	0.4024	2.44671	< 0.01
negative tone	-0.843236	0.3508	0.503793	0.044
afraid	-0.976545	0.7637	3.15626	< 0.01
amused	-1.03066	0.7445	3.05194	< 0.01
angry	-0.904103	0.7876	3.1977	< 0.01
annoyed	-0.957057	0.7703	3.13157	< 0.01
don't care	-1.00875	0.7524	3.13915	< 0.01
happy	-0.991061	0.7587	3.15777	< 0.01
inspired	-0.951042	0.7724	3.13281	< 0.01
sad	-1.27825	0.642	3.13403	< 0.01

ADF stands for Augmented Dickey-Fuller test for log-return

Source: author's computations.

Table A.4: Stationarity tests-hourly logarithmic differenced values

Log-returns	ADF	P-value	KPSS	P-value
happiness	-18.2425	8.545e-044	0.0101989	> 0.1
caring	-17.5991	2.867e-042	0.0104871	> 0.1
depression	-19.0528	1.463e-045	0.0095168	> 0.1
inadequateness	-21.136	3.164e-049	0.0145229	> 0.1
fear	-18.9837	2.036e-045	0.0246464	> 0.1
confusion	-21.4967	1.008e-049	0.0142972	> 0.1
hurt	-18.2115	1.006e-043	0.0110793	> 0.1
anger	-20.5242	2.746e-048	0.00367525	> 0.1
loneliness	-17.3829	9.84e-042	0.00890268	> 0.1
remorse	-20.2419	8.152e-048	0.0127122	> 0.1
positive mood	-18.2839	6.872e-044	0.00511568	> 0.1
negative mood	-19.1625	8.71e-046	0.0107615	> 0.1
positive tone	-18.4424	3.015e-044	0.0131129	> 0.1
neutral tone	-30.03	2.177e-048	0.0281349	> 0.1
negative tone	-18.938	2.538e-045	0.0207883	> 0.1
afraid	-17.3747	1.032e-041	0.00481663	> 0.1
amused	-17.0114	8.696e-041	0.00313532	> 0.1
angry	-17.3188	1.426e-041	0.00336392	> 0.1
annoyed	-17.2165	2.588e-041	0.00309071	> 0.1
don't care	-17.1744	3.313e-041	0.00295814	> 0.1
happy	-17.2217	2.51e-041	0.00307737	> 0.1
inspired	-17.1574	3.661e-041	0.0029584	> 0.1
sad	-16.9737	1.089e-040	0.00295228	> 0.1

ADF stands for Augmented Dickey-Fuller test for log-return

Source: author's computations.

Table A.5: Stationarity tests

Log-returns	ADF	P-value	KPSS	P-value	ADF-GLS	P-value
JY	-37.093	0.0000	0.219218	>0.10	-3.16011	0.001541
SAP	-10.8882	4.43e-022	0.226017	>0.10	-2.49662	0.01214
SV	-11.1015	8.95e-023	0.139628	>0.10	-3.64255	0.0001
CD	-10.3312	2.85e-020	0.0739586	>0.10	-5.30524	1.68e-007
CL	-18.7105	7.75e-045	0.028627	>0.10	-8.33684	6.07e-015
NG	-22.0464	2.13e-050	0.128737	>0.10	-4.37264	1.33e-005

ADF stands for Augmented Dickey-Fuller test for log-return, while ADF-GLS for Augmented Dickey-Fuller GLS test for log-return

Source: author's computations.

Table A.6: Additional test statistics related to the explanatory power of the HAR-RV model

	Plain model	happiness	Caring	Depression	Inadequateness	Fear
LLR-test		0.00008133	0.00004913	0.0000572	3.119e-06	0.04349975
MSE	6.82e-12	6.45e-12	6.44e-12	6.44e-12	6.32e-12	6.69e-12
DM statistic		1.629	1.59	1.66	1.977	1.002
P-value		0.1033	0.1119	0.0969	0.0481	0.3164

Source: author's computations.

Table A.7: Additional test statistics related to the explanatory power of the HAR-RV model

	confusion	hurt	anger	loneliness	remorse	afraid
LLR-test	0.00001213	7.346e-07	0.0081887	0.00037597	0.00144304	1
MSE	6.39e-12	6.53e-12	6.63e-12	6.51e-12	6.56e-12	6.51e-12
DM statistic	1.445	2.179	2.112	1.601	1.302	1.626
P-value	0.1484	0.0293	0.0347	0.1094	0.1928	0.1040

Source: author's computations.

Table A.8: Additional test statistics related to the explanatory power of the HAR-RV model

	mood positive	mood negative	tone positive	tone neutral	tone negative	tag Nasdaq
LLR-test	0.00001487	0.00012603	0.00110269	0.00016134	0.0005553	0.00032457
MSE	6.39e-12	6.47e-12	6.55e-12	6.48e-12	6.53e-12	6.66e-12
DM statistic	1.192	1.328	1.203	1.371	1.259	1.895
P-value	0.2333	0.1843	0.2291	0.1703	0.2080	0.0581

Source: author's computations.